ANALYTIC ALTERNATIVES FOR EVALUATING HUMAN SERVICES INTERVENTIONS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Nicholas Phillips Huntington

January 2013

ANALYTIC ALTERNATIVES FOR EVALUATING HUMAN SERVICES INTERVENTIONS

Nicholas Phillips Huntington, Ph.D.

Cornell University 2013

Human service programs and clients are complex. Clients in publicly-funded human service interventions frequently are dealing with multiple, complex, co-occurring issues and conditions, and human service agencies are increasingly responding with comprehensive, integrated, individualized service approaches.

The goal of this dissertation is to investigate the utility of alternative analytic techniques that might provide greater "representational complexity" — richness and detail of information about clients and their interactions with programs — than standard, "off-the-shelf" analytic techniques provide. Two analytic approaches that are not often used in human service evaluation studies but that might lead to higher levels of representational complexity are assessed in two separate sub-studies. Within each sub-study the same evaluation data is analyzed using an alternative technique and a current, standard technique.

The first study, a comparison of effectiveness between two programs, draws on data from a national cross-site evaluation of substance use treatment for youth. In this study the alternative technique Maximum Change Scores (MCS — Boothroyd, Banks, Evans, Greenbaum, & Brown 2004) is tested against the standard technique for this situation, Hierarchical Linear Models (HLM — Raudenbush & Bryk, 2002; Snijders & Bosker, 2012; Singer & Willett, 2003).

The second study, an examination of within-program patterns of differential effectiveness, draws on data from a case management program for homeless women.

In this study, the tree-based modeling technique Classification and Regression Trees (CART — Breiman, Friedman, Olshen, & Stone, 1993) is pitted against standard linear models for examining whether different types of clients showed different patterns of change while in the program.

The innovative techniques are evaluated against their standard counterparts on five dimensions: 1) comparability of results, 2) representational complexity, 3) utility to service providers, 4) validity, and 5) data requirements.

In the first study, MCS produced results that were overall comparable with those from HLM. The technique worked as expected in representing clients change more flexibly, but its usefulness is limited because it is not inherently longitudinal and is less flexible than HLM. In the second study, the tree models produced results that differed substantially from the linear model results but provided higher representational complexity.

BIOGRAPHICAL SKETCH

Nick Huntington received a BA in Psychology from Harvard University in 1987 and a MA from Cornell's Department of Human Development in 1995. Since the late 1990's Nick has been working in human services research and evaluation. He is interested in improving research and evaluation methods for better understanding human service programs and clients.

ACKNOWLEDGEMENTS

This dissertation was a long time in the making. I thank all the friends, family members, teachers, administrators, and colleagues who have helped and supported me in completing this work over the decades and across two millennia.

The deans and staff at the Graduate School were extremely flexible, forgiving, and supportive in allowing me to finish after such a long period. Dean Ellen Gainor, with great humanity and mercy, made the original call to allow me to continue, and the final determination was supported by Dean Tilman Baumstark, for which I am grateful. The two GSR's I worked with, Janine Brace and Mike Skinner, answered endless questions and were wonderful in facilitating all the administrative arrangements that needed to be made.

In Human Development the GFR's Barbara Koslowski and then Elaine Wethington were also enthusiastic supporters of my finishing, for which I am grateful. Throughout, Bonnie Biata supported the work from the Department's end.

Besides being a friend and supporter, Elliott Smith acquired administrative targets and dispatched them with ease during my A-Exams, which was extremely helpful.

My mom provided finances to help support my family while I worked on the dissertation, for which I am grateful.

Along the way friends including Elliott, Alison, Jai, Steve, Bill, Mary Ann, Jack, Susan, Brian, Ellen, Lisa, Lee, Deb, and Jasper provided support, encouragement, coffee, and, once, brisket.

Wendy Vaulton and Cheryl Amey, as colleagues and friends, have been a huge source of both academic help and psychological support through dark times and light along the way.

At Advocates for Human Potential, my boss Amy Salomon gave me the flexibility, support, space, and encouragement to complete this project. Without her kind mentoring and help, and the approval of AHP leadership, it never would have happened. Terri Tobin, Jen Carpenter, Laura Elwyn, David Centerbar, and Denise Lang all took on extra work and shouldered extra burdens while I was focusing on this project instead of doing my job. Their encouragement was a big source of motivation.

My Special Committee: John Eckenrode and Mon Cochran, who swapped as chairs, and Bill Trochim, provided a never-ending stream of positivity, good cheer, and encouragement, as well as insights, feedback, ideas, and inspiration. They are models for me of academics engaged in real-world issues and practice. They could have dropped out at any time over the last fifteen years, but they stuck with me, which was a huge support.

My wife Noelle and my children Max, Eliza, and Anna put up with a lot of crankiness, absence, and occasional despondency over the years that I worked on this. Noelle tirelessly took on my family responsibilities and made lunches I should have made, drove kids to lessons when I should have, and even made the ultimate sacrifice of taking the family to Disney without me so I could focus on school. Their flexibility, encouragement, and love were critical ingredients in this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Human services programs and clients are complex. The clients in publicly-funded human service interventions frequently are dealing with multiple, complex, co-occurring issues and conditions (Minkoff, 2001; Kessler, Chiu, Demler, Merikangas, & Walters, 2005; U. S. Department of Housing and Urban Development, 2011). Human service agencies are increasingly responding with comprehensive, integrated, individualized service approaches (Drake, O'Neal, & Wallach, 2008; McHugo et al., 2006) focusing on the whole person, recovery, and wellness (Substance Abuse and Mental Health Services Administration, 2011; Kaplan, 2008; Swarbrick, 2006).

How can researchers trying to understand human services programs address this complexity? In the face of such complexity, the relative rigidity of traditional quantitative research methods used in the behavioral sciences can seem inadequate to the task. McGuire and Mcguire (1988) note that behavioral research often has a "reverse-Midas touch" and turns interesting and important topics into "dross" by "cutting them to fit a Procrustean bed of conventional methods" (p. 97).

The goal of this dissertation is to investigate the potential utility of two alternative analytic techniques for addressing the complexity inherent in human services evaluations.

In this Introduction chapter I first briefly describe the two evaluation contexts in which the techniques might be useful and then proceed to the main thread of argument for why I think this investigation is important. In those sections I lay out some concepts motivating the current work and then for each of the two evaluation contexts review the current methodological approach and the alternative technique I will examine. The Introduction concludes with a statement of the

study's framework and research questions.

1.1 Evaluation Research Contexts

Before proceeding it is important to define the scope of the evaluation contexts being considered. Just as the objects of study in human services evaluation (the clients, programs, and contexts in which they operate) are complex, so too is human services evaluation itself. Evaluation scholars have varying definitions of evaluation and taxonomies of types of evaluation. Rossi and Freeman (1985) describe three classes of evaluation research: analysis related to program conceptualization and design, to program monitoring, and to assessment of program utility. Mark, Henry, and Julnes (2000) have a similar but slightly different taxonomy outlining four goals of evaluation: assessment of merit and worth, program and organizational improvement, oversight and compliance, and knowledge development. Both of these conceptualizations contain the distinction between formative and summative evaluation, first proposed by Scriven in the 1960's (see Shadish, Cook, & Leviton, 1991, p. 74).

In this dissertation I will focus on two important evaluation contexts, one summative and one formative. The first, summative, research context is when there are two programs or treatments to be compared and the question is "Is program or treatment A effective in comparison with program or treatment B?". This is the prototypical summative evaluation question, and is often of interest to policy makers and funders. Sophisticated research projects addressing this question might involve random assignment of clients to the two treatments being compared, but in everyday evaluation practice in the human services random assignment is rare. The second context is a formative one, and concerns situations in which there is a single program and the practitioners of the program are interested in answering the question: "Within our program, do particular sub-groups of clients show more

or less change than others, are we being differentially effective with different types of clients?". This question is often of more interest to those more deeply involved in the day-to-day operations of programs such as program directors or front-line service providers. Service providers will frequently have ideas about the types of clients they have, or have not, been able to help, based on their own (potentially somewhat anecdotal) experiences.

Taking these two contexts as central ones in human services evaluation, this dissertation will examine the potential utility of two alternative analytic methods in answering the cross-program effectiveness and within-program differential effectiveness questions. In doing so, I will effectively conduct two sub-studies, one for each evaluation context, and within each study compare the current best-practice methodology to an alternative technique.

1.2 Causal Granularity of Outcomes and Representational Complexity

In thinking about the complexity of human service programs and the methods we have to study them, it may be helpful to think in terms of two concepts: "causal granularity" and "representational complexity", that underlie how I have thought about this issue and motivate this work.

The first concept, causal granularity, derives from discussions of the philosophy of evaluation by Michael Scriven. Scriven has been described by evaluators as "our philosopher" and, along with Donald Campbell, is seen as one of the foundational theorists in the philosophy and methods of evaluation (Shadish et al., 1991). In his work on the philosophy of social science, he contends that the social sciences are "unlucky" because the things we care about explaining in the social sciences

are much more complex than those that we care about explaining in the physical sciences. In his view, "the difference between the scientific study of behavior and that of physical phenomena is thus partly due to the relatively greater complexity of the "*simplest phenomena we are concerned to account for* in a behavioral theory" (Scriven, 1956 reproduced in Martin & McIntyre, 1994, p.72, emphasis in original). He gives a helpful analogy for this difference in complexity between what the physical and social sciences must explain. He notes that physics can accurately predict what will happen when something falls in a vacuum, but is less able to predict what will happen when an object falls through air, and that "when it comes to the question of how a particular leaf falls from a particular tree on a particular autumn day, we are almost helpless" (Scriven, 1964, p.171 cited in McIntyre, 1994). Fortunately for physics, those are not the sort of problems which physicists are tasked with addressing but, unfortunately for behavioral scientists, those are the sort of problems that are interesting and worth addressing in our arena.

He elaborates on the basic problem the social sciences face in saying that, compared to the physical sciences,

> The basic generalizations are more complex, in the sense that more standing conditions must be specified for a functional relationship of comparable simplicity, consequently more variables must be measured in obtaining the basic data to which the basic generalizations refer. (Scriven, 1956, p. 72)

This idea is very close to what I am calling "causal granularity of outcomes". Scriven is saying that the social sciences are more complex because you need to specify more variables to adequately capture the state of something, and therefore be able to relate that state to a future state or to the state of something else. If we

think in terms of a program's potential impacts, "causal granularity" is how ecologically narrow the factors are that drive patterns in program outcomes. Scriven's notion of many variables being required to describe something would correspond to small causal granularity, that is that the causes that drive outcomes are specific rather than broad. At the high or "broad" end of the scale, one can imagine a relatively simple program with a clearcut goal, such as a driver's ed program where the outcome is simply passing the driver's test. This type of program would have relatively high causal granularity, because the causal factors that shape outcomes are not very idiosyncratic, numerous, interacting, limited in time and space, and specific to particular persons. At the other end might be programs like substance abuse treatment programs where it seems that factors shaping outcomes could be very numerous and context-dependent and that patterns of outcomes might therefore be very finely structured. Substance abuse outcomes might be importantly impacted by a myriad of factors such as the brain chemistry of the client, the nature of the client's addiction, their attitudes and values, how much social support they have, the fit between the substance abuse providers and the client, the nature of other issues the client is dealing with, the service system in which the program is embedded, the current local cost and availability of illegal substances, and numerous other factors.

In their excellent work grounding evaluation in a realist philosophy of science, Pawson and Tilley (1997) describe the logic of explanation as follows:

> The basic task of social inquiry is to explain interesting, puzzling, socially significant regularities (R). Explanation takes the form of positing some underlying mechanism (M) which generates the regularity and thus consists of propositions about how the interplay between structure and agency has constituted the regularity. Within realist investigation

> there is also investigation of how the workings of such mechanisms are contingent and conditional, and thus only fired in particular local, historical or institutional contexts (C). (Pawson & Tilley, 1997, p. 71)

Using their concepts, we could define causal granularity as the degree to which mechanisms (M) are contingent and conditional on the context (C). Presumably, this quality varies for different mechanisms in different contexts, and some mechanisms are more broadly applicable, while others are narrower and more limited in the scope of contexts in which they operate.

If we assume hypothetically that two exactly identical people who received exactly the same services in a program must have the same outcomes from the program, then causal granularity indexes the extent to which the characteristics of the two hypothetical people and their contexts and services received can be different from one another before we are likely to see differences in outcome. Does the slightest minute difference in, say, brain chemistry between the two participants lead to different outcomes, or can they be relatively different on many factors before we are likely to see different outcomes?

The notion of causal granularity of outcomes is clearly related to the notion of validity in outcome evaluation, which has been a topic of lively debate among evaluators (Chen, Donaldson, & Mark, 2011; Shadish, 2011). Campbell and co-authors (Campbell, 1957; Campbell & Stanley, 1963; Cook & Campbell, 1979) originally discussed validity for outcome evaluations in the context of describing research designs for these studies and the threats to valid inference to which the different designs are susceptible. They originally outlined two types of validity: *internal* validity which is whether the treatment impacted outcomes in the particular program under study, and *external* validity which indexes the extent to which the causal relationship can be generalized to other populations, settings, and outcomes. Cook

and Campbell (1979) further refined each of these two types producing a four-category typology of statistical conclusion validity, internal validity, construct validity, and external validity. Campbell subsequently (Campbell, 1986) advocated relabeling internal validity to "local molar causal validity", which addresses the question "did this complex treatment package make a real difference in this unique application at this particular place and time?" (p. 69). The term *molar* is meant to emphasize a recognition that the treatment is often a "hodgepodge", perhaps involving hundreds of different potentially important causal elements. In the same paper Campbell advocates abandoning the strict term external validity for a more graduated *principle of proximal similarity*, which is that we should expect the outcomes in a different study to be the same to the extent that the situation is similar in terms of populations, context, and other factors. It appears this is really the same concept as causal granularity as I have thought of it and Scriven's notion of complexity, but oriented more towards comparisons between programs rather than delving down within the causal effects of a single program (that is within the "local molar" causal effect). In any case, all of these conceptualizations have at their root the idea that outcomes should be more similar for people and contexts that are more similar, and less so for comparisons across people and contexts that are more dissimilar.

The second concept, what I am calling "representational complexity", is thankfully perhaps more easily defined. I see this as a property of methodologies. It is the degree of complexity, richness, and detail that the method provides in representing findings about people and how they interact with programs. All methodologies provide some simplified representation of reality. This concept indicates simply how complex vs. simple that representation is. Methodologies with higher representational complexity allow a more detailed and specific representation of

the interplay of person, process, and context (Bronfenbrenner, 1979) or regularity, mechanism, and context (Pawson & Tilley, 1997), than methods with lower representational complexity. One can imagine methodologies arrayed on a continuum on this construct. At one end might be qualitative case studies of clients' experiences in a program. These case studies could provide extremely nuanced and detailed findings concerning the interplay of multiple factors, for example how specific characteristics, attitudes, or perceptions of particular clients impacted their interaction with specific program staff members, and how those interactions fostered, or failed to foster particular outcomes. At the other end of the continuum might be simple group-level statistical tests that simply provide a "yes/no" answer indicating whether clients in one group had, on average, higher or lower scores on an outcome than clients in a comparison condition.

These two ideas of causal granularity and representational complexity are linked in that one is a characteristic of people and programs, indexing how complex they are, and the other is a characteristic of methodologies, and how complex a representation they allow. Presumably, in a situation of broad causal granularity, methodologies of low representational complexity would be adequate. But, in situations where the mechanisms are very local and detailed, richer methodologies would be needed to not inadvertently abstract over important differences. Presumably, human services interventions tend towards this latter class, and tend to have complicated, local, narrowly defined causal granularity where client outcomes are impacted by a wide array of interacting factors. These interventions would therefore require high levels of representational complexity in the methods used for their analysis so that we can adequately represent their complexity.

In the remaining sections I will, for each of the two evaluation contexts, describe the standard analytic approach used and the proposed alternative method.

I argue the proposed methods have the potential to provide higher representational complexity than the standard methods allow.

1.3 Cross-Program Examinations of Program Effectiveness

Evaluating the effectiveness of different programs, treatments, or conditions by quantitatively comparing the outcomes for clients enrolled in each has a long history and is a dominant, mainstream activity in human services evaluation. The current approach in the broader conceptual and methodological sense derives from the tradition of social experimentation and causal effects as developed by methodologists including Campbell (Campbell & Stanley, 1963; Cook & Campbell, 1979), Holland (Holland, 1986), Scriven (Scriven, 2008), and Rubin (Rubin, 1974, 2005). Although the "causal wars are still raging, and the amount of collateral damage is increasing" (Scriven, 2008, p. 11), we will not delve into the major current controversy concerning whether randomized controlled trials should be the only acceptable form of evidence for interventions. Rather, since RCT's are so rare in human services, we will focus instead on the theory and justification of non-randomized intervention studies. Campbell produced the foundations in describing "quasi-experiments" (non-randomized experiments), the potential threats to validity that must be guarded against in interpreting quasi-experiments such as history, selection, and maturation, and the strengths of different quasi-experimental designs (Campbell & Stanley, 1963; Cook & Campbell, 1979).

Working from a more statistical background, Rubin (Rubin, 1974, 2005) and Holland (Holland, 1986) developed the "counterfactual" theory of program effects, which provides a framework for thinking about causal effects in intervention programs and the appropriate data analytic approaches for these situations. The novel and crucial idea here is that what we would ideally like to estimate in such situa-

tions are the outcomes for a set of units (e.g. clients) in the treatment condition, and the outcomes *for the same units* in the control condition, so that we could then calculate a measure of relative program impact within each person (e.g. by taking each person's outcome under the treatment condition minus their outcome under the control condition). These quantities would represent the true causal impact of the program. We could then average across clients to get a measure of the average impact of the program. Sadly, we are unable to have the same people go through both conditions, so for each participant we have an actual condition that they experienced (i.e. they were a treatment client or a control client), and a "counterfactual" condition that they did not experience (the opposite condition). Since we cannot have the same client, in their exact current situation, both go through and not go through a program, we estimate these unknowable causal effects by comparing the outcomes for clients who did go through the program to those of a surrogate group of people who are as similar as possible to the clients but did not go through the program (a comparison group). This theory thus lays out the logic for why comparing outcomes to a comparison group approximates an important quantity that we wish to know.

1.3.1 Hierarchical Linear Models

Turning now from the conceptual foundations to practical analytic approaches, how would one analyze client outcome data to estimate a program's causal effect? Over the past 20 years one technique, hierarchical linear models (HLM—Raudenbush & Bryk, 2002; Snijders & Bosker, 2012; Singer & Willett, 2003) has become the standard method both for longitudinal analysis in non-intervention settings and for estimating the counterfactual causal program effects described above (Fitzmaurice, Laird, & Ware, 2004; Gelman & Hill, 2007). Historically, hierarchi-

cal linear models came about through the confluence of two bodies of work: that of social scientists concerned with understanding the role of context and addressing multiple, nested levels of analysis, and that of statisticians who developed *mixed effects* models in which some effects are assumed to be fixed (the familiar type of effect from statistical modeling), and some are allowed to be random, that is to vary randomly from subject to subject. Researchers realized that both contextual variability and individual variability could be examined in a regression modeling framework by including a random effect of context (e.g. neighborhoods) and a nested random effect for subjects (e.g. people within neighborhoods) (Snijders & Bosker, 2012). The popularity of hierarchical models (also known as multilevel models, mixed effects models, or random effects models) in evaluation was greatly enhanced by Raudenbush and Bryk's (2002) book which made them more accessible and described them mostly in educational evaluation where the effect of context — students in the context of classrooms in the context of schools in the context of school districts — is particularly salient and important.

In longitudinal data analysis with hierarchical models, besides whatever larger ecological structures are included, there is an additional level of "hierarchy" of time points nested within people. At the bottom level of the hierarchy, the model fits to each person's data the line (or other curve if desired) that best fits their individual scores on the outcome over time. Conceptually, each subject therefore has an individual-specific "intercept", representing where they started out on the outcome, and an individual-specific "slope", representing their rate of change over time on the outcome (slope in the mathematical sense of "rise over run", how much the outcome for this particular client changed per unit time). The intercept (starting place) and slope (rate of change) can be modeled as a function of other factors. The other factors may be characteristics of the individual, measures of

their involvement in the program, and, if there are multiple programs being compared, characteristics of the programs as a whole. When a human services study involves clients in different programs, therefore, each individual's change on an outcome (their slope) is typically modeled as a function of program-level factors (e.g. the size of the providing agency) in addition to person level factors (e.g. severity of substance abuse history), and service measures (e.g. how many counseling sessions they participated in). Additional levels of nesting are possible, for example if the intervention involves one-on-one counseling and measures about counselors are available (e.g. their degree of professional certification), this additional level of predictor variables could be added.

With HLM analysis, then, each client is allowed to have their own specific rate of change, some people may have large slopes indicating a big improvement on a measure, say a measure of quality of life, others may have slopes near 0 indicating no substantive change in quality of life, and others may have negative slopes, indicating a decline in quality of life. The analysis can uncover factors that predict higher or lower levels of change for clients. So one might determine for example, that clients who had less severe mental health issues entering the program improved more in quality of life than those who entered with more severe issues. One can also model whether clients in one program or treatment have higher or lower rates of change than clients in another program or treatment, controlling for other factors, thus providing a direct test of program effectiveness. This method clearly allows for a much higher level of "representational complexity", a much richer description of the complex interaction between clients and program, than can be accomplished with simple statistical techniques. HLM, however, is subject to the same restrictions of all linear models (such as multiple regression). The number of independent variables that can be included is limited, depending on

the number of cases, and interactive effects, where the predictive power of one independent variable depends on the value of another, are difficult to handle. The impacts of predictors on clients' rates of change are actually included into the model by specifying interactions with time (e.g. gender by time), but in practice identifying more complex interactive patterns with multiple variables is unlikely.

1.3.2 Alternative Technique: Maximum Change Scores

As described above, a methodologically sophisticated human services evaluation examining the efficacy of a program would likely use a non-randomized control group and hierarchical models to estimate treatment effects, perhaps also with propensity scoring, a technique to improve inferences by better controlling for pre-treatment differences in non-randomized groups (c.f. Rosenbaum & Rubin, 1984).

Boothroyd, Banks, Evans, Greenbaum, and Brown (2004) added to this toolkit a method called maximum change scores (MCS) that might be useful in increasing the "representational complexity" available in these research contexts. They developed this method explicitly to increase the flexibility with which analysts can approach multi-faceted human service programs in which services might be multiple and individually tailored. The approach is very straightforward. Using their procedure, the analyst creates an outcome variable that is the maximum standardized change score of each client across a set of outcome measures. Within each client, this variable represents the change from pre- to post-treatment on the outcome for which they experienced the most positive change. This change score, even though it represents different outcomes for different people, is then analyzed as the dependent measure. Thus, the program is allowed to benefit different clients differently. For example one client may improve the most on a measure of mental health status, while another may improve the most on a measure of substance abuse

status, or quality of life. In their use of the method Boothroyd et al. (2004) were analyzing data from a randomized experiment, and thus they tested for between-program differences using only follow-up scores. In the non-randomized context, I will be following their procedure but also modeling the change score to help control for baseline differences.

1.4 Within-Program Examinations of Differential Effectiveness

Program administrators, front-line service providers, and evaluators working in human service programs often wish to know whether their program is helping the full range of clients they serve, or whether instead there are sub-groups of clients that are not being adequately helped. Frequently, service providers have hunches based on their day-to-day experience that clients with particular demographic or clinical characteristics are, or are not, doing as well as they might hope (Pawson & Tilley, 1997). Adamson, Sellman, and Frampton (2009) note that if practitioners have an accurate understanding of the factors that predict outcomes they can improve their practice in at least three ways: by identifying client groups that are not being helped, by identifying specific changeable factors they can impact that in turn may impact outcomes, and improving the accuracy of prognosis for clients, thus allowing for improved treatment planning.

Several recent reviewers in human service fields have commented on our inability to predict which clients will do well, and not do well, with particular treatments. Uher (2011) highlights the role of two factors in producing this situation: the rise of standardized and widely adopted diagnostic criteria and the rise of evidenced-based practice. He notes that these forces have led to a predominant clinical approach in

which human services practitioners diagnose clients based on the constellation of the client's symptoms and conditions, and then select a treatment approach based on the diagnosis or condition. Within this reasonable framework, however, he notes how practitioners are generally unable to predict differential program effectiveness:

> with the primary role of evidence-based medicine fulfilled, shortcomings are becoming apparent, and many perceptive clinicians wonder if something has been lost along the way. The principal weakness of evidence-based medicine is that it does not account for the vast variability in therapeutic response within each diagnostic group. (Uher, 2011, pp. 109–110)

In a comprehensive review of the effectiveness of treatments for people with co-occurring mental health and substance abuse disorders, Drake et al. (2008) similarly highlight "interventions for subgroups" as an issue needing future work. They note that "in all intervention studies, dual diagnosis clients respond variably to a particular intervention or program. If diagnosis is not a strong predictor of treatment response, perhaps we should search for other ways of identifying subgroups for future intervention studies" (Drake et al., 2008, p. 135). In the introduction to a special issue in the *Harvard Review of Psychiatry* on the topic, Roffman (2011) makes the same point, lamenting "our inability to predict confidently which patient will respond to which treatment" (Roffman, 2011, p. 99).

Conceptually, when a client is engaging in a formal human service program of some type there is a potentially vast array of factors that could shape treatment outcomes including characteristics of the client, the service providers, the program, the setting, the larger services environment, the involvement of other collateral people to the client such as family or friends, the client's motivation and expectations, the fit between the client and the services provided, the timing and

sequencing of treatment components, and the level of engagement and retention that is achieved along the way towards more distal hoped-for program outcomes. In the field of mental health Norcross and Wampold (2011) note that researchers have identified over 200 client characteristics alone that might potentially impact treatment outcomes, over 100 of which have been investigated empirically.

1.4.1 Linear Models of Pre/Post Data

From the point of view of practitioners, the inability of the field to identify generalizable predictive factors is probably of secondary importance compared to understanding patterns of differential effectiveness *here and now in their particular program.* Likewise, evaluators working with human services programs often wish to provide real-time feedback on differential effectiveness, based on systematically collected information, thereby hopefully bypassing any preconceptions or biases that service providers may be operating under in their more anecdotal or experience-based understanding. For program evaluators who take seriously a formative approach to evaluation, in which part of the job of a program evaluator is to help providers improve their practice, answering such questions is a central concern.

A variety of analytic approaches are feasible when examining differential effectiveness within programs, including the hierarchical models described above, but in this context we will focus on the simpler situation where only pre- and post-data is available, and forgo discussion of data with multiple timepoints and more sophisticated longitudinal modeling. For practitioners wishing to understand differential effectiveness in their own programs, frequently pre/post data is all that is available, for example from intake and exit assessments.

The central analytic problem here is determining whether clients with different characteristics have different outcomes, while controlling for baseline status.

This seemingly simple problem, of estimating the effect of factor X on a post-test outcome Y, while controlling for the pre-test values of Y, has a long history of debate in the statistical literature, apparently stretching back to the 1960's (Allison, 1990). Besides several secondary methods, there are three main ways such data can be analyzed; the factor (or factors) of interest can be related to: 1) the change score (the post-score minus the pre-score), or 2) the residuals left over after regressing the post-score on the pre-score, and thus removing the linear effect of the pre-score on the post-score, or 3) the post-score while controlling for the pre-score by including it in the statistical model (Bonate, 2000; Twisk & Proper, 2004; Allison, 1990). Twisk and Proper analyze the same data using all three techniques and conclude that the third method gives the best results. This method is often called an analysis of covariance (ANCOVA) model because it is a linear model with an often categorical factor (the X) and an often continuous covariate (the Y pre-test). In a follow-on to Twisk and Proper's work, Forbes and Carlin (2005) concur in finding the ANCOVA model the best approach. In his book-length treatment of the subject involving simulation studies, Bonate (2000) also finds the ANCOVA model superior. All three authors find support for the ANCOVA approach largely on the grounds that it does the best job at controlling for "regression to the mean", the statistical artifact by which change can be related to pre-test values, even if no causal mechanism is at play.

A reasonable approach for the evaluator seeking to understand patterns of differential effectiveness would therefore be to model outcomes as a function of client background characteristics while controlling for the pre-test values of the outcomes. Like HLM described above, this approach has limitations common to all linear modeling techniques such as a limitation on the number of variables that can be included, independence and distributional assumptions, and difficulty in finding

and making explicable higher-order interactions among predicting variables.

1.4.2 Alternative Technique: Tree-Based Methods

To potentially provide a higher level of "representational complexity" than the linear (ANCOVA) models allow, I explore the use of classification and regression trees, or CART (Breiman, Friedman, Olshen, & Stone, 1993; Strobl, Malley, & Tutz, 2009) in this evaluation context of examining differential effectiveness. CART comes from the rapidly growing fields of *data mining* and *machine learning* (Witten & Frank, 2005; Ripley, 1996). The focus of these fields is on

> techniques for finding and describing structural patterns in data as a tool for helping to explain that data and make predictions from it (Witten & Frank, 2005, p. 9).

CART is one of the most widely used of these pattern finding techniques. It is a quantitative classification methodology in which, as in standard linear regression, the analyst seeks to understand variability in a dependent measure as a function of a set of independent or predictor variables. The algorithm used however is quite different from that of standard linear models. In CART the predictor variables can be any mix of categorical and continuous measures and the dependent variable can also be categorical or continuous. Although different types of CART exist, most work by *binary recursive partitioning*. Under this algorithm, all of the cases start in a single group. The computer examines each of the predictor variables and finds the variable, and the particular value of the variable, that were the cases to be split into two groups based on that cut point, most cleanly divides the cases in terms of their values on the dependent variable. That split of the cases into two groups is made and the algorithm proceeds by repeating the procedure

within each group formed by the first split. For each "branch", the variable and splitting value that best splits the cases in the group in terms of the dependent variable is found. In this way an upside-down "tree" of successive binary splits is made. Groups at the bottom of the tree, defined by their values of the predictors involved in the splits, have varying levels of the dependent variable. Note that this method is not statistical in the sense of postulating a sample and a population, estimating parameters, and making inferences based on distributions. It is simply mathematical, finding the splits in variables that maximize the difference between the resulting groups on the outcome variable.

In the original CART algorithm (Breiman et al., 1993), the splitting process continues until the "leaves" of the tree are homogeneous with respect to the dependent variable or contain only one case. Such tree models have been criticized for "over-fitting" the data, and being extremely sensitive to minute particularities of the dataset under analysis (Strobl et al., 2009). One approach to guard against over-fitting is to "prune" the tree through a reverse procedure of removing the least helpful splits. Hothorn, Hornik, and Zeileis (2006) have recently developed a statistically motivated version of CART, called *conditional inference trees* that uses statistical criteria to only conduct splits that result in statistically significant reductions in the "impurity" (i.e. dependent variable heterogeneity) in the resulting groups.

CART produces, then, a tree-like structure of splitting rules based on the independent variables. Each rule is of a form such as "Gender = Male" or "Height < 5'-9"" and divides cases into two smaller groups. This set of rules ends up classifying cases in terms of their likely value on the dependent variable. The rules can be used to predict a new case's likely value on the dependent variable and, more importantly for our purposes, they can be used to explicate the factors related to

the dependent variable. CART may be useful because:

- It is inductive, that is it searchers for patterns without their being determined *a-priori*
- It can draw upon a large number of predictors without the problems that arise when trying to estimate too many parameters in a linear model,
- The predictors and the dependent variable are not restricted to follow any particular statistical distribution, and
- Perhaps most importantly, it very naturally handles complex interactions that could not be handled through standard linear models with interaction terms.

This beneficial handling of complex, interactive patterns in the data results from the hierarchical nature of the splitting process. To illustrate, a CART analysis of factors predicting heart attack risk might reveal that the first most powerful split is on gender with, say, women being at lower risk than men. At the second level of the tree, the variable that best splits the men into two groups based on their risk might be weight while the variable that best splits the women might be another characteristic, such as age. The CART analysis has therefore revealed an interaction in predicting heart attack risk in which the importance of two variables (weight vs. age) depends on gender. The ability to handle complex patterns in predictors holds the best promise for this technique enhancing the "representational complexity" over the standard linear model approach. Conceivably, the CART methods could uncover rich, descriptive patterns in outcomes that the linear models cannot because the required interaction terms would be too difficult to estimate and interpret in a linear modeling framework.

1.4.3 Alternative Technique: Random Forests

Breiman (2001) developed the technique of *random forests*, which are ensembles of individual CART models. The logic of ensemble methods is to increase predictive accuracy, and decrease sensitivity to particularities in the data, by producing many models on subsets of the data and then averaging over them (Strobl et al., 2009). In random forests the individual CART models produced and averaged over are based on subsets of the data in two ways. First, each tree is based on a randomly chosen "bootstrap" sample, that is a sample drawn with replacement from the original data, so that some units are represented more than once and some are not represented at all. Second, only a random subset of the predictor variables is allowed into each tree model, which introduces another large source of variability. By making the component trees highly diverse, and then combining the predictions from them, the forest approach increases the robustness of the results. A grave drawback to the forests however, is that unlike single trees in which one can see the actual tree model produced, there is no corresponding easily interpretable output from a forest analysis. Because the forest is based on many diverse trees, there is no simple output that shows the results. The methods do however produce measures of variable importance across, of relatively how important each variable was in predicting the outcome, when averaged across all the trees of the forest (see Methods). Forests increase predictive accuracy but sadly do not have an easily interpretable representation.

1.5 The Current Study: Framework, Goals, and Specific Research Questions

This dissertation will explore the utility of these two alternative quantitative techniques: maximum change scores for the cross-program effectiveness question and tree-models for the within-program differential effectiveness question. The hope is that these techniques will increase the representational complexity we can bring to bear in studying complex clients interacting with complex programs.

The basic approach will be to conduct two studies, one for the cross-program effectiveness context, and one for the within-program differential effectiveness context, and in each study analyze the same data twice, once using the standard "off-the-shelf" technique, and once using the innovative, non-standard technique, and contrast and compare what we learn from the different approaches. Therefore, four analyses will be conducted. The two studies, and the methods being compared under each, are:

- Study A: Program Effectiveness - Is program or treatment A effective in comparison with program or treatment B? This question is often of interest to policy makers and researchers. The comparison here is between:
  - Hierarchical linear models
  - Maximum Change Score methods
- Study B: Within-Program Differential Effectiveness - Within a program, do particular sub-groups of clients show more or less change than others? This question is often of interest to program directors and front-line service providers. The analytic techniques to be assessed here are:
  - Pre/post linear models

– Tree-based data mining methods, specifically conditional inference trees and random forests

The alternative techniques under examination have several characteristics that make them of theoretical interest. First, and most importantly, they each might increase the representational complexity at our disposal and broaden the range of possible findings in ways that can make them more subtle, individualized, and nuanced than the findings that are possible from traditional techniques. Second, neither of the techniques is deeply or purely statistical. While they are both quantitative analysis techniques, they each involve changes in how we conceptualize the research problem inherent in outcome studies, and thus engage the issue further back in the research process than at the final analysis stage. Such reconceptualization may be necessary if the techniques we use in evaluation studies are ever to do justice to the complexity of their subjects. Third, both techniques are readily adoptable by everyday program evaluators in real-world contexts. Both are relatively simple to implement and neither requires sophisticated statistical knowledge or specialized software. Finally, the two techniques might have the potential to work well together, and to play off each other's strengths in such a way as to make the use of both of them simultaneously stronger than the use of either individually.

This study can be seen as fitting into (and hopefully contributing to) long-running streams of methodological developments for the two different evaluation contexts being considered. This work has the potential to extend each, and potentially help add a useful tool to the set of methods that have been developed for each of these evaluation contexts. In both cases, there has been movement towards more sophisticated methods allowing for more representational complexity, and this study can be seen as trying to further those progressions. In the case of cross-program effectiveness, the methods have a long history tracing back to Fisher's

original work on experimentation and randomization (see Campbell, 1986). Since the methods inherently involve change over time, their development is intimately tied up with the development of statistical methods for handling longitudinal data. Apparently, the first analysis of longitudinal data can be traced to Student's use of a t-test on pre/post scores in 1908 (Hedeker & Gibbons, 2006). The analytic methods have developed through four broad phases (Hedeker & Gibbons, 2006; Fitzmaurice et al., 2004), the first of which was the use of simple statistical tests on summary measures that collapse across time, as with Student's t-test of change scores. The second broad approach was repeated measures Analysis of Variance (ANOVA) which allows for subjects to have different starting points but not different trajectories over time, and makes very unrealistic assumptions about the data in a longitudinal context. The third phase was repeated measures Multivariate ANOVA (MANOVA) which allows the analyst to test for overall group effects and for polynomial changes in time (e.g. linear, quadratic) but imposes the restriction that all subjects have data from all timepoints. Finally, hierarchical linear models were developed and applied to longitudinal data analysis, greatly increasing the flexibility and potential representational complexity since subjects are allowed to vary in both starting place and trajectory, variation in these values can be modeled explicitly, and missing data is allowed. The maximum change score approach potentially extends this progression by allowing clients in a program to be differently affected by the program, and show change on different outcome measures.

While not as developed as those for the cross-program case, the methods for within-program examinations of differential effectiveness have also progressed from simple statistical tests on summary scores (for example via *post-hoc* t-tests of change scores by a covariate of interest such as gender) to statistical modeling via ANCOVA, which explicitly controls for pre-treatment levels of the outcome

(Bonate, 2000). As discussed above, while such modeling is more effective, it suffers from the serious limitation of all linear models, that it is very difficult to find and interpret interactions, especially higher-order ones. The tree methods examined here have the potential to increase the flexibility and representational complexity we have at our disposal in such work, by being useable with both categorical and continuous outcomes, handling any number and mix of predictors, inductively finding structure in the data, and most importantly deftly handling interactions, which are likely the rule rather than the exception given the complexity of the programs and clients under study.

In examining the utility of these alternative techniques in these two different evalution contexts, I will address five specific questions:

1. To what extent are the results from the standard and innovative techniques comparable? Logically, it seems that four outcomes are possible:
    - The standard and innovative approaches produce largely similar results; the innovative techniques do not materially add to the conclusions we are able to draw from the data using the standard techniques.
    - The standard and innovative approaches produce different but conformable results; the innovative techniques might for example provide more detailed results that fit within general results produced by the standard techniques.
    - The standard and innovative approaches produce different results that are unrelated to each other. The two sets of results do not contradict each other but are not interlinked or supportive of one another either. They are simply different.
    - The standard and innovative approaches produce contradictory results. Findings from the innovative techniques in some way argue against con-

clusions one reaches when using standard techniques.

2. Do the innovative techniques provide for a more complete and richer representation of the (presumed) complexity of people and programs than the standard techniques allow? Do they give us higher representational complexity than the standard methods? This question will only be relevant under certain outcomes of the first question.
3. Is one set of results (innovative vs. standard) more useful to service providers involved in the program than the other? (This question will only be addressed for the differential effectiveness question).
   (a) Is one set of results more easily understandable to providers who lack specialized research training?
   (b) Is either set of results inappropriate in terms of scope, providing too much detail or too little, too "forest" or too "trees", from the providers' point of view?
   (c) Does one set of results have more "face validity" than the other set? Does one conform more closely to service providers' views of the program and its impacts for clients?
   (d) Is one set more useful for improving program practices? It seems possible that different approaches may lead to types of results that are more or less useful in this regard.
4. Do the innovative techniques provide results that are as solid, reliable, and defendable as the conclusions that are drawn from the standard techniques?
5. Do the innovative techniques make higher demands of the data than do the standard techniques, for example requiring more subjects for comparable levels of statistical power or confidence?

The next two chapters contain the methods and results for Study A and Study

B, while in the final chapter I discuss the findings across both studies, and relate what was found using the different techniques to the above primary study questions.

CHAPTER 2

STUDY A: CROSS-PROGRAM EXAMINATION OF PROGRAM EFFECTIVENESS

Study A focuses on between-program comparisons of effectiveness and draws upon data from the Assertive Adolescent and Family Treatment (AAFT) evaluation.

2.1 Methods

2.1.1 Program and Data

The data for Study B derives from the Assertive Adolescent and Family Treatment (AAFT) Cross-Site Evaluation. The AAFT program was funded in 2006 by the Substance Abuse and Mental Health Services Administration of HHS (SAMHSA) to help implement and disseminate evidence-based treatment for substance abuse in youth. Under this ongoing grant program, grantees are funded to implement a specific evidence-based treatment model, the Adolescent Community Reinforcement Approach (A-CRA) followed by Assertive Continuing Care (ACC) (Godley et al., 2001). This model emphasizes involvement of family members and others in adolescents' lives, and consists of a highly manualized set of different procedures that counselors can draw upon in their work with youth. The A-CRA phase of the treatment, which is intended to last approximately 12 weeks, is followed by ongoing case-management help for clients in the Assertive Continuing Care phase of the model. The model is notable for its extensive training and implementation supports. The creators of the model run an ongoing training and certification program and clinicians seeking certification need to attend an extensive in-person training and submit digital session recordings for fidelity assessment.

The A-CRA/ACC model has been shown to be as effective as other models for helping youth deal with addiction. In the federally-funded Cannabis Youth Treatment (CYT) study, one of the largest studies of substance abuse treatment for youth conducted to date, the A-CRA/ACC model was compared in a randomized trial with four other evidence based models and found, in common with the others, to lead to increased days of abstinence at 12 months following enrollment, and an increase in the percentage of youth in recovery. A-CRA/ACC was also found to be one of the most cost-effective of the models studied (Dennis et al., 2004).

In 2009 SAMHSA funded a cross-site evaluation of 14 agencies that received funding that year under the AAFT program. This cohort of AAFT grantees was the first in which agencies had a choice of serving either adolescents, defined as age 12 to 17, or transition-age youth (TAY), defined as ages 18 to 24. As a condition of receiving funding, grantee agencies were required to collect client-level data through in-person interviews with clients at baseline, 3 months, 6 months, and 12 months post-enrollment. Data collection could be done by clinicians but in most cases was done by a separate staff member, often from an external evaluation agency. The data collection was standardized, all grantees were required to use the Global Assessment of Individual Needs (GAIN), a comprehensive bio-psychosocial structured assessment tool (Dennis, Titus, White, Unsicker, & Hodgkins, 2003).

I chose two sites for analysis from the full cohort of 14 agencies that were funded and participated in the cross-site evaluation. Since the goal of these analyses is to compare the two methods (HLM and MCS) for addressing the basic question of program effectiveness, I wanted a "clean" comparison of two sites, rather than the very much more complex and frequently harder to interpret comparison of multiple sites. The two agencies were chosen because 1) both served transition-age youth, 2) neither stood out as serving an extremely different population than

other AAFT sites, and 3) they differed somewhat in the overall strength of their implementation, as assessed by the cross-site evaluation. Site A had a mid-range implementation strength and Site B had a lower implementation strength. Site A is located in a southern city and Site B is located in a small New England city and includes the surrounding rural area.

2.1.2 Measures

The GAIN is a large instrument with eight core sections: Background, Substance Use, Physical Health, Risk Behaviors and Disease Prevention, Mental and Emotional Health, Environment and Living Situation, Legal, and Vocational. It produces over 100 scales that can be used for diagnosis, level of care placement, and treatment planning according to established clinical standards (Ives, Funk, Ihnes, Feeney, & Dennis, 2010). For research and evaluation purposes, scales from the GAIN can be used to assess lifetime severity of issues in various domains and for measuring change. Some scales are broad indices with relatively low internal consistency, some are more focused scales that assess longer-term status and can be used as predictors or covariates, and some assess immediate or recent timeframes and can be used as outcomes. In selecting variables for inclusion in this study, I aimed to represent a range of different domains both for the outcomes and for the measures of background conditions. Where possible I obtained both a background measure and an outcome measure from a given domain. Table 2.1 shows the GAIN scales selected.

I chose to examine outcomes in three domains: Substance use, mental health, and physical health. Substance use was the main area of focus for the AAFT grantees and the defining criterion for clients to be admitted into the projects. Mental health issues frequently co-occur and interact with substance use problems

Domain	Type	Name/Description
Substance Use	Background	Lifetime Substance Problems Scale ($\alpha = .90$) – Recency with which the respondent experienced symptoms of substance abuse, dependence, and substance abuse induced psychological disorders based on DSM-IV.
Substance Use	Outcome	Substance Frequency Scale ($\alpha = .80$) – Avg of percent of days in past 90 that respondent used alcohol or other drugs. Includes items assessing days experienced heavy use and days experienced problems from use.
Mental Health	Background	Internal Mental Distress Scale ($\alpha = .94$) – Count of symptoms experienced in the past year related to internalizing disorders (e.g. depression, anxiety).
Mental Health	Background	Behavioral Complexity Scale ($\alpha = .94$) – Cont of symptoms experienced in the past year related to externalizing disorders (e.g. attention deficit, hyperactivity, conduct disorder).
Mental Health	Outcome	Emotional Problems Scale ($\alpha = .79$) – Recency (during past 90 days) and number of days bothered by, and kept from responsibilities by, emotional problems.
Physical Health	Outcome	Health Problems Scale ($\alpha = .73$) – Recency (during past 90 days) and number of days bothered by, or kept from responsibilities by, physical/medical problems.
Trauma	Background	General Victimization Scale ($\alpha = .82$) – Count of types of victimization experienced and traumagenic factors involved in victimization.
Criminal Behav.	Background	General Crime Scale – Number of different types of illegal activities engaged in in past year.

Table 2.1: GAIN Scales

(see Introduction) and have also in my experience been most amenable to relatively short-term interventions such as the AAFT program. Physical health was chosen as a domain that was less closely inter-twined with the other two, but nevertheless important for overall well-being.

2.1.3 Hierarchical Linear Models

As explained in the Introduction, hierarchical linear models (HLM) have become the standard in longitudinal data analysis, and since examining program efficacy almost always involves assessing clients before and after exposure to some type of intervention, they have become the standard technique used in relatively sophisticated human services program evaluations. When used in a longitudinal context, the "hierarchy" of the hierarchical model is of assessment points nested within clients, who are themselves possibly nested within agencies, clinics, programs, etc. In the context of a human services evaluation, hierarchical models can be used to examine change over time, and differential change over time for clients in different programs, while controlling for baseline differences.

The basic framework for these models is to predict the outcome from three sets of predictor (independent) variables: demographics (age, gender, and race), severity of conditions, and site (the indicator for being in Site A vs. Site B). For the severity variables, those variables for domains other than the outcome are included in each model, so for example, the model for the mental health outcome includes the background condition measures for substance abuse, trauma, and criminal behavior while the model for substance use includes the mental health, trauma, and criminality measures as predictors.

For each of the three outcome measures, I fit a series of five multilevel models, starting with the "unconditional" model as a base. This general approach is one recommended by Singer and Willett (2003) and Bickel (2007). The unconditional model includes no predictors and simply partitions the variation in the outcome into a "within-person" component (how much a client's scores vary around that client's mean score) and a "between-person" component (the variation of each client's overall mean to the grand mean across clients). This is very similar to the

partitioning of variance in an analysis of variance.

The second model in the series then adds time as a predictor at level one of the model, that is within-person. This allows a separate slanted line to be fit to each person's set of up to four data points (the assessments at baseline, 3, 6, and 12 months). The slope of this line therefore represents the person's change in the outcome variable over time. This time effect is entered into the model as both a fixed and a random effect so that the model estimates the *overall slope* of the line across clients, but also allows each client to have their *own unique slope* over time.

This parameterization of the model allows us to examine whether background factors impact clients' starting places (their baseline values of the outcome variable), and their rates of change over time (their slopes). In the statistical model, the effects for rate of change are tested by entering "cross-level interaction terms" between a background characteristic, say gender, and time. For gender, this term would test whether the effect of time, that is how the outcome changes over time (the slope of the outcome over time), differs between men and women.

The third model in the series adds site as a predictor, therefor testing whether clients at one site had higher or lower starting levels of the outcome at baseline and whether they had more or less change over time than clients at the other site. The fourth and fifth models add demographics and background severity of conditions as predictors.

These models, complex as they are, are straightforward compared to the range of models that can be fit using the technique. While it would conceptually make sense to fit three-level HLM models, it was not possible in this case. In a three-level model, timepoints would be considered nested within clients and clients nested within sites. With only two sites, it is not possible to estimate the third level and site in these analyses is treated as a fixed, person-level covariate akin to other non-

varying aspects of a person such as age or gender. Furthermore, other time-varying covariates could be included so that change in one outcome might be modeled as a function of change in another, or as a function of (presumably increasing) levels of services received over time. Also, it is possible to treat time more flexibly than as a linear trend. In the models fit here each client has their own starting place and rate of change over time. It is possible to fit curvilinear time effects or to transform time or the outcome so that the outcome changes in a linear fashion with time. I did not use time-varying covariates or curvilinear effects for time simply to keep the overall complexity of the models manageable and to facilitate straightforward comparisons with the MCS technique.

2.1.4 Maximum Change Scores

The maximum change scores analysis follows the method described by Boothroyd et al. (2004). To make the analysis comparable with the HLM models, I used the same three outcome variables in forming the change score. The maximum change score is calculated by first calculating standardized change scores on the three outcomes of interest by dividing each client's change from baseline to follow-up by the standard deviation of the measure at baseline. In order to form a single score for analysis, I set each client's follow-up score to their latest follow-up point, whether it was 3 months, 6 months, or 12 months. These scores thus represent how much each person changed during their time in the program relative to the amount of variability found in the measure between people at baseline. These scores can be positive or negative, depending on whether the client increased or decreased on the measure between timepoints. Since the three outcomes are all measures of problems, where a higher score indicates more severe impairment, having a negative change score indicates an improvement in status over time.

The second step in forming the MCS is to pick, within each subject, the standardized score on which they experienced the most beneficial change. For each client, this was their *smallest* change score from among the three calculated. For most clients this is a negative number, indicating that they improved from baseline to follow-up, but for some it is positive, indicating that a client's best score was their smallest worsening in condition between baseline and follow-up.

While these "raw" maximum change scores can be examined, Boothroyd et al. (2004) recommend transforming this score to enhance its statistical properties by using a cumulative normal distribution so that the final scores range between 0 and 1. This is the score used for final analysis with standard distribution-based statistics. They also use a non-parametric transformation of the raw score, taking the ranks, and use that outcome measure as a non-parametric alternative to be analyzed with non-parametric methods.

The intervention which Boothroyd et al. (2004) examined using this method was a randomized trial of home-based services for children. Because their study was randomized, it was sufficient to compare the programs on the maximum change score using simple statistical tests. Since human service evaluations, like the AAFT project, are rarely randomized, I examine the change score using linear models controlling for the same demographic and background factors as used in the HLM analysis. Thus, the independent and dependent variables entering both analyses are comparable.

2.2 Results

After profiling the AAFT study participants and their changes on key outcome measures, I present the results of analyzing the AAFT data using both the standard HLM methodology and the alternative Maximum Change Score approach.

2.2.1 Profile of AAFT Participants' Background

While the two AAFT sites both served nominally similar populations, that is both targeted "transition-age-youth (TAY)" with substance use issues, in practice there can be great variability in populations enrolled in different agencies, especially in different cities with varying demographics and service systems. In this case, Site A is located in a large urban area in Florida while Site B is in a small city in New England.

Table 2.2 shows the demographic characteristics of youth enrolled at the two AAFT sites. Site B enrolled a larger proportion of women than Site A but, unsurprisingly given that it was a small New England city with surrounding rural counties, was substantially less diverse racially. Both sites served transition age youth, mostly in the 18-20 range.

		Count(Pct)	
Measure		**Site A**	**Site B**
Female		15 (24.6%)	27 (39.1%)
Age			
	15-17	5 (8.1%)	2 (2.9%)
	18-20	45 (73.8%)	52 (75.4%)
	21-29	11 (18.0%)	15 (21.7%)
Race			
	African American	16 (26.2%)	2 (2.9%)
	White	26 (42.6%)	57 (82.6%)
	Hispanic	12 (19.7%)	1 (1.4%)
	Mixed	6 (9.8%)	6 (8.7%)
	Other	1 (1.6%)	3 (4.3%)

Table 2.2: Demographic Characteristics of AAFT Participants by Site

As described in Methods above, the GAIN assessment tool used by AAFT sites contains a broad range of scales assessing the severity of various issues and conditions. Figure 2.1 shows the distributions of six key measures within each of the two sites. In the boxplots, the box covers the 25^{th} to 75^{th} percentiles. The

first measure, the Substance Problems Scale, is a count of symptoms of substance abuse, substance dependence, and health and psychological problems associated with use. The next two measures assess mental/emotional health, the first internalizing symptoms and the second externalizing symptoms associated with attention deficit, hyperactivity, and conduct disorders. The fourth scale is an index of types of traumatic events experienced, and their traumagenic characteristics. The fifth scale indexes types of criminal activity the respondent has engaged in, and the final scale summarizes information from a very broad range of subscales included in the GAIN tool, and is thus partly duplicative with the information in the other scales. Participants across the two sites are broadly similar on these measures, with the important exception of the substance abuse measure. Participants in Site B have substantially higher substance abuse issues. The mean at Site A is 8.8, compared to 11.1 at Site B (out of a total of 16 possible symptoms), so the difference in averages between the sites is more than two symptoms. This difference is statistically significant ($t(124) = 2.77$, $p=.006$). Probably because of this strong difference, the sites approach having a statistically significant difference on the overall cross-domain measure as well ($p=.09$).

Compared to Site A, Site B is more diverse in terms of gender, less diverse in terms of race, and is serving a population with more severe substance use issues.

2.2.2 Change on Outcome Measures for AAFT Participants

The distributions of the three key outcome measures: substance use frequency, emotional problems, and health problems, are shown in Figures 2.2, 2.3, and 2.4. These plots are descriptive, simply showing the distribution at each site and assessment point without modeling, and without any subsetting based on patterns

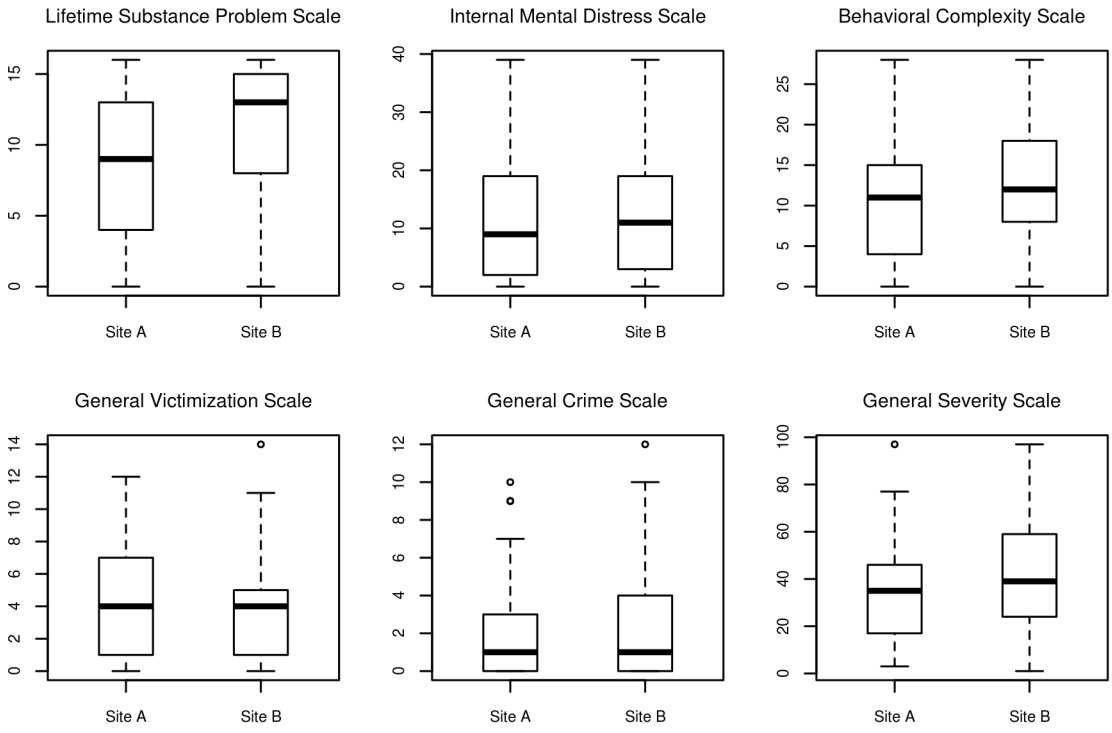


Figure 2.1: Severity of Background Conditions Among AAFT Participants

of missing data.

The three outcomes appear to show different patterns of change, when examined at the group level. On the substance use measure (Figure 2.2), the pattern appears different between the two sites. Paralleling the difference between the sites in severity of clients' lifetime substance abuse problems noted above, the baseline value of the immediate substance frequency measure here – based on days of use in the past 30 – is notably higher for participants at Site B than at Site A. Participants at both sites appear to have relatively low levels of use in the follow-up waves. As is frequently the case when measuring severity of issues, the distributions are generally left-skewed (here bottom-skewed) with many clients at the low end of the scale and a long tail of higher values indicating a handful of clients with more extreme conditions.

For the mental health measure, clients appear to have elevated levels of emo-

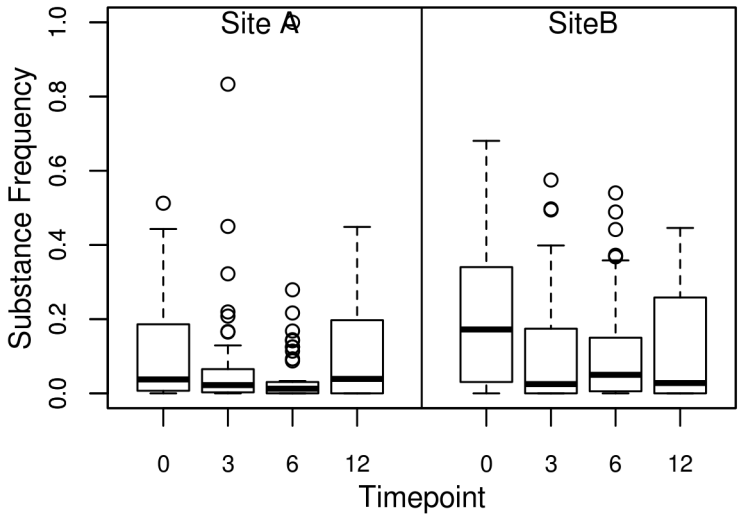


Figure 2.2: Distribution of Substance Use Frequency by Site and Assessment Point

tional problems at baseline compared to the follow-up waves, and the pattern is quite similar across the two sites.

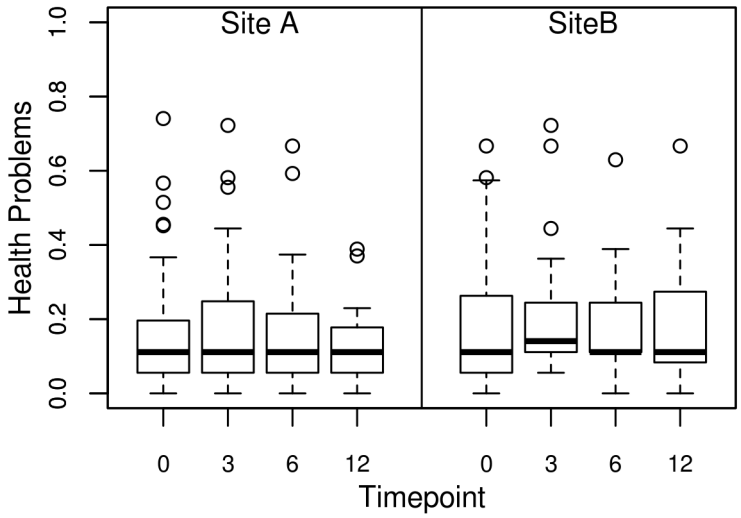


Figure 2.3: Distribution of Emotional Problems Score by Site and Assessment Point

For the health problems measure, there appears to be little change over time and also little difference between the sites.

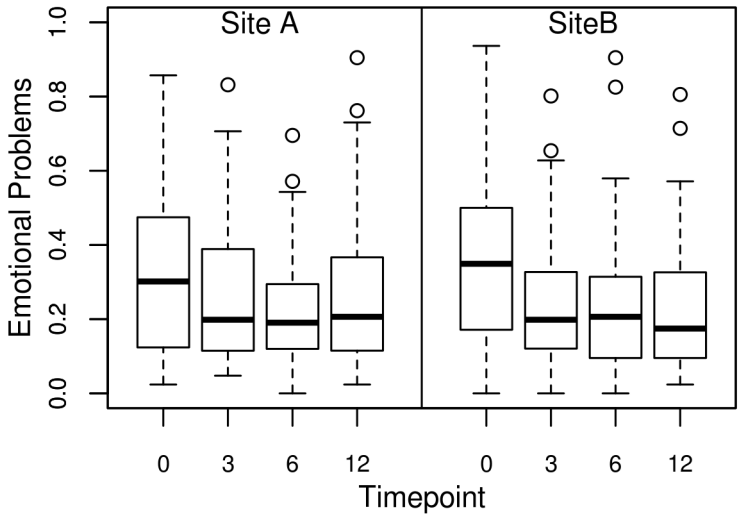


Figure 2.4: Distribution of Health Problems Score by Site and Assessment Point

As expected, the substance use and mental health outcomes appear more intertwined with one-another. These two are significantly correlated at the 12-month follow-up ($r(53) = .46$, $p < .001$). The health measure is not significantly correlated with either the mental health or substance use measure.

2.2.3 Hierarchical Linear Models Analysis

This section contains the results of fitting the series of hierarchical linear models (see Methods) to each of the three outcome variables: substance use frequency, emotional problems, and health problems. In reporting these models, I take the explicitly multilevel approach used by Singer and Willett (2003) and Raudenbush and Bryk (2002) over more of a "mixed effects" presentation used by more statistically inclined authors such as Snijders and Bosker (2012) and Crawley (2002). Using this framework, the effects in each model are described as having an impact on initial status – in this case clients' level of the outcome variable at baseline – and having an impact on change over time – in this case clients' change per

month on the outcomes between baseline and follow-up assessments. In the tables that follow, under the fixed effects section, the impact of predictors on baseline status is reported first, and the impact of the predictors on rate of change is underneath. Note that the effects of the predictors on rate of change are actually entered into the model as a predictor-by-time "cross level" interaction terms (see Methods above). Thus, for example, the effect of site on rate of change is technically entered into the model as a site-by-time interaction term. Those familiar with longitudinal modeling may be more familiar with this way of describing the effects. The bottom portion of each table reports the random variance components of the model and the model's overall fit to the data.

Table 2.3 shows the series of five models fit for the first outcome, substance use frequency. The first model fit, the Unconstrained model, establishes the baseline against which the other models are compared. This model contains no predictors, just an overall mean for substance use frequency at baseline (the intercept term under Baseline Status), and it helpfully partitions the variation in the outcome into the within-person, and between-person components, reported under the Variance Components section of the table. The model indicates that the predicted level of use at baseline is 11.9 (on the 100-point scale), close to the overall mean obtained without modeling, which is 11.6. This this value is significantly different from 0, meaning we can conclude that there is substance use in the sample at baseline. The within-person variance represents the variability of clients' data points around their own personal mean. The large number here relative to its standard error (140.6/12.5 = 11.2), means that there is a significant amount of variability within subjects and that systematic change over time such as a downward trend, or some other external factors driving why people have particular levels at particular timepoints are likely at work. This high number means that

there is within-person variability in substance use frequency that is likely "worth explaining". The between-person variability is of roughly the same scale as the within-person variability and is also large compared to its standard error. This represents the variability of clients' baseline substance use scores around the mean of the scores across clients. This relatively high number indicates that the differences between clients are also likely "worth explaining" by adding covariates. The proportion of the total variability in client baseline scores that is between-subject, the *intraclass correlation*, is 44%.

The second model in the table, labeled Time, adds time as both a fixed and random effect to the base Unconditional model. The Rate of Change portion of the table shows how the predictors relate to the "slope" for time, that is the rate of change per month in the outcome. The figure for the intercept in this portion of the table is -.40. This is *not* the intercept for the outcome at baseline (which is above and is 13.4 in this model), but rather the basic predicted value, without other covariates, of *the change in the outcome per month.* This figure, which is statistically significant (p=.016), indicates that clients' substance use scores declined by about .4 per month, so by the end of 12 months we can expect a decline of about 3.6 points. Adding the Time effect to the Unconditional model reduces the within-person variability from 140.6 to 129.6, indicating that fitting a slanting line to each client's timepoints reduces the within-person error compared to simply assuming that their timepoints are flat at the level of their personal mean, which the Unconditional model does.

This Time model, because it includes both fixed and random effects for time, allows each client to have their own unique slope for time. Including only a fixed effect for time would allow clients' levels to change linearly with time (rather than being flat) but would constrain all the clients to have the same slope. The

		Unconditional	Time	Time + Site	Demographics	Conditions
Fixed Effects		Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)
Baseline Status	Intercept	11.9 (1.1) ***	13.4 (1.4) ***	8.8 (1.9) ***	10.8 (2.6) ***	10.0 (1.7)
	Site			8.7 (2.6) **	7.6 (2.9) **	6.5 (2.3) **
	Age 21 or older				3.3 (3.2)	
	Female				-2.5 (2.8)	
	Nonwhite				-3.6 (3.0)	
	Internal Mental Distress					.48 (.16) **
	Externalizing Problems					.57 (.20) **
	Victimization					-1.2 (.44) **
	Crime					.44 (.46)
Rate of Change	Intercept		-.40 (.17) *	-.05 (.23)	-.17 (.32)	-.06 (.22)
	Site			-.68 (.33) *	-.71 (.36) *	-.60 (.31)
	Age 21 or older				-.22 (.43)	
	Female				.44 (.35)	
	Nonwhite				.10 (.36)	
	Internal Mental Distress					-.05 (.02) *
	Externalizing Problems					-.05 (.03)
	Victimization					.17 (.06) **
	Crime					.03 (.05)
Variance Components		Variance (SE)	Variance (SE)	Variance (SE)	Variance (SE)	Variance (SE)
	Within-person	140.6 (12.5)	129.6 (14.2)	128.8 (14.2)	128.9 (14.2)	129.8 (14.2)
	In Baseline Status	112.5 (21.1)	147.2 (30.5)	128.8 (28.2	124.1 (27.6)	82.3 (22.8)
	In Rate of Change		.45 (.45)	.37 (.43)	.33 (.43)	.05 (.38)
Model Fit						
	Deviance	3150.8	3141.4	3130.5	3126.3	3091.6
	AIC	3156.8	3153.4	3146.5	3154.3	3123.6
	BIC	3168.7	3177.1	3178.1	3209.7	3186.8
	LR Test Chi-square (df)		9.4(3) *	10.9 (2) **	4.2 (6)	38.9 (8) ***

Table 2.3: Multilevel Models for Substance Frequency Scale

variation between clients in the slope appears small compared to its standard error (.45 compared to .45), suggesting that it is not significantly different from 0, and perhaps all clients can be considered to have the same slope and a random effect is not needed. When looked at compared to the value of the average slope itself however (-.40), the variance of .45 seems large. Multiple statisticians and HLM experts have discussed the difficulties in interpreting variance components and in particular the unreliability of determining their importance by comparing them to their standard errors (e.g. Singer & Willett, 2003; Rabe-Hesketh & Skrondal, 2012; Mitchell, 2012), so we will retain the random effect in the subsequent models.

The fit statistics at the bottom of the table generally indicate that the Time model is an improvement over the Unconditional model. The deviance, which represents the "lack of fit" of the model compared to a perfectly fitting one, is lower. Of the two "adjusted" deviance measures, the AIC is lower as well, while the other, the BIC, which applies a more stringent penalty for adding parameters, increases. The final row of the table shows the likelihood ratio chisquare test of each model to the logically preceding one, testing whether the reduction in deviance between models is statistically significant.

The third model, labeled Time+Site, adds site to the basic Time model. We find that site is important both for the initial baseline value ($p < .01$) and for the change per month ($p < .05$). With site in the model, the overall estimated baseline value is lower (8.8), but the coefficient for site, which represents the additional effect on starting value of a client being in Site B, is quite large and significant. In fact, the model estimates that clients in Site B start with scores approximately double those of clients in Site A, which parallels the descriptive results above. In terms of change over time, now that site has been added to the model we find that the overall rate of change per month is smaller (-.05) and no longer significant.

The coefficient for site for the rate of change portion of the table represents the additional change that clients in Site B experience, relative to those in Site A. The change for clients in Site B is estimated at -.73 per month (the -.05 overall value plus the additional -.68 due to being in Site B). The initial estimate of a decline of .40 per month overall in the Time model has been decomposed into estimates of .05 per month for those in Site A and .73 per month for those in Site B. The reduction in deviance between the Time + Site model and the Time model is significant, indicating better fit.

The fourth and fifth models add two sets of predictors, demographics and background conditions, to the basic Time + Site model. The Demographics model does not change any of the previous coefficients in qualitatively different ways and none of the demographics have an impact on either the starting value or the rate of change. Both of the adjusted deviance criteria and the likelihood ratio test concur in indicating that nothing is gained by adding demographics to the model.

Since no new predictors entered in the Demographics model, the Conditions model builds off the previous Time + Site model and adds in four background conditions: Internal mental distress, externalizing problems, extent of victimization, and involvement in crime. Here we find that a one point increase in the internalizing and externalizing problems scales is associated with a .48 and .57 increase respectively in starting level of substance use. Interestingly, clients scoring higher on victimization have lower levels of initial use. Involvement in criminal activities has no relationship to starting level. Looking at rate of change, it appears that victimization is associated with an increase in use, while internal mental distress is associated with a decline in use, perhaps because those with higher levels of internal distress might have higher levels of baseline substance use, and therefore more room to decline over time. Controlling for these background factors makes

the statistically significant site effect disappear.

Table 2.4 has the corresponding set of models for the emotional problems outcome. The Unconditional model, as with the substance use measure, shows considerable variability within subjects around their own personal mean, and between subjects in their baseline levels. The intraclass correlation for this outcome is in a middle range, .33, meaning that a third of the variability in starting level is between subjects.

When time is entered in, in the second model, the overall rate of change is -1.0 per month, which is statistically significant (p=.036), indicating that clients experience improvements, though small, in mental health status over time. The within-person variability also declines significantly, now that time is in the model, and the model fit criteria are unanimous in favoring the Time model over the Unconditional model. Even the very strict BIC criteria is somewhat lower, and the likelihood ratio test indicates a strongly significant improvement in fit.

There appears to be little difference between the sites. Adding site in the third model results in the overall time effect decreasing in magnitude to -.74 per month, but staying statistically significant. In this model, that value of -.74 represents the change per month for clients in Site A (the base category for site). Clients in Site B are not significantly different in either starting place or rate of change. So to this point these models indicate there is a slight decline in emotional problems over time, and that there is no difference between the sites in this effect.

In the Demographics model, the only significant effect is for female on the baseline value. Women are estimated to have starting scores 7.5 points higher than men. Addition of the demographics makes the coefficient for the intercept in the rate of change portion of the model non-significant (it drops from .74 in the previous model to .62, p=.20). This coefficient for the intercept has to be

		Unconditional	Time	Time + Site	Demographics	Conditions
Fixed Effects		Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)
Baseline Status	Intercept	27.7 (1.4)**	31.2 (1.7)***	29.9 (2.4)***	26.8 (3.3)***	26.1 (2.9) ***
	Site			2.3 (3.3)	1.8 (3.6)	.58 (3.3)
	Age 21 or older				2.5 (4.1)	.56 (3.7)
	Female				7.5 (3.5)*	4.7 (3.3)
	Nonwhite				1.4 (3.7)	1.1 (3.4)
	Substance Use Problems					1.1 (.34) **
	Victimization					1.7 (.49) **
	Crime					-.2 (.57)
Rate of Change	Intercept		-1.0 (.25)***	-.74 (.35)*	-.62 (.48)	-.61 (.48)
	Site			-.45 (.49)	-.55 (.53)	-.60 (.52)
	Age 21 or older				-.68 (.64)	-.66 (.63)
	Female				.35 (.53)	.49 (.55)
	Nonwhite				-.13 (.54)	.05 (.55)
	Substance Use Problems					-.01 (.06)
	Victimization					-.01 (.08)
	Crime					.06 (.09)
Variance Components		Variance (SE)	Variance (SE)	Variance (SE)	Variance (SE)	Variance (SE)
	Within-person	279.3 (24.6)	226.2 (25.1)	226.6 (25.2)	228.5 (25.4)	237.3 (26.2)
	In Baseline Status	136.1 (30.0)	193.8 (47.1)	191.7 (46.9)	176.5 (45.4)	103.7 (37.7)
	In Rate of Change		1.8 (1.0)	1.7 (1.0)	1.6 (1.0)	1.3 (.95)
Model Fit						
	Deviance	3338.6	3319.2	3318.4	3307.2	3257.5
	AIC	3345.5	3331.3	3334.4	3335.2	3297.5
	BIC	3357.4	3354.9	3365.9	3390.4	3376.4
	LR Test Chi-square (df)		20.3 (3)***	0.9 (2)	11.1 (6)	49.7(6) ***

Table 2.4: Multilevel Models for Emotional Problems Scale

interpreted in the context of the other variables in the model. It represents the amount of change per month when the other covariates are at their "base" values. The base values for categorical covariates are their excluded reference categories, and the base values for any continuous covariates are their means (though there are no continuous covariates in this model). So in this model, the intercept in the rate of change portion of the table represents the decline per month for white males in Site A who are less than 21 (the combination of the base categories for the other covariates). This decline is not significant and none of the other factors (being at Site B, being female, being nonwhite, or being older) make for a rate of change that is significantly different from this value. This raises an interesting issue in interpretation. It is also the case that one can test in a post-hoc manner the significance of time in other sub-groups. For example if you do a post hoc test of the change over time for white males who are less than 21 in Site B (instead of those in Site A), the value is a statistically significant decline of 1.2 per month. This raises issues around how to parameterize models and the difficulties in exploring all the possible effects that could occur, as the number of combinations of predictors is enormous (please see the Discussion section for more on this issue).

The overall fit of this model is not better than the Time+Site model. The BIC criterion is especially critical, penalizing this model for adding many parameters but yielding only a small improvement in fit, and the likelihood ratio test is only marginally significant (p=.08). Because of the clinically meaningful difference of 8 points in starting status for females however, I elect to retain this model as the basis for the final model.

In the final model, I add the background conditions to the set of predictors from the Demographics model. I retain all the predictors from the Demographics model, even though almost all are non-significant, because the comparison of models via

the deviance criteria can only be accomplished with models that are proper subsets/supersets of one-another, so one is always testing a simpler model against a more complex model that includes all the terms of the simpler one, plus additions. Background conditions are predictive of baseline levels of emotional problems, but not of change over time. Youth with higher levels of substance use problems at baseline have higher levels of emotional problems, and youth with higher levels of traumatic experiences have higher levels of emotional problems. Neither of these findings is surprising, given the frequent co-occurrence of mental health issues, substance use, and trauma. Interestingly, however, change in emotional problems over time was not significantly different by any of the background factors. This final model fit the data notably better than the Demographics model: Both the deviance criteria decline and the likelihood test is highly significant.

The parallel set of models for the health problems scores is presented in Table 2.5. The first, Unconditional model is similar to that for the previous two outcomes in illustrating a large degree of intra- and inter-individual variability and a intraclass correlation in the mid range (here .38). The Time model appears to improve the fit, but the fixed term for the rate of change is not significant (.07 with a standard error of .17). This indicates that the fit is improved because of the random time effect, allowing each person to have their own time slope, rather than because there is a common time trajectory across clients.

The addition of site in the third model makes no substantive difference; clients at both sites appear to start in roughly the same place and have roughly the same rate of change on the health measure. There is some hint of an effect of site on rate of change, because the overall (intercept) coefficient for the rate of change becomes more negative and the coefficient for being in Site B is positive, but the coefficients are all small and non-significant.

		Unconditional	Time	Time + Site	Demographics	Conditions
Fixed Effects		Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)	Coeff. (SE)
Baseline Status	Intercept	16.9 (1.0)***	17.1 (1.3)***	16.6 (1.8) ***	17.1 (2.5) ***	16.3 (2.4) ***
	Site			.98 (2.6)	-.95 (2.8)	-2.4 (2.6)
	Age 21 or older				3.5 (3.2)	2.1 (3.0)
	Female				3.5 (2.8)	2.9 (2.7)
	Nonwhite				-3.5 (2.9)	-2.7 (2.7)
	Internalizing					-.05 (.17)
	Externalizing					.38 (.20)
	Substance Use					1.0 (.31) **
	Victimization					.26 (.45)
	Crime					-1.1 (.48) *
Rate of Change	Intercept		-.07 (.17)	-.15 (.24)	-.28 (.33)	-.11 (.32)
	Site			.17 (.34)	.37 (.37)	.51 (.35)
	Age 21 or older				-1.0 (.45) *	-.87 (.43) *
	Female				.00 (.36)	-.13 (.39)
	Nonwhite				.50 (.37)	.31 (.37)
	Internalizing					.01 (.03)
	Externalizing					-.04 (.03)
	Substance Use					-.10 (.04) *
	Victimization					.05 (.06)
	Crime					.08 (.07)
Variance Components		Variance (SE)	Variance (SE)	Variance (SE)	Variance (SE)	Variance (SE)
	Within-person	130.6 (12.0)	114.8 (13.0)	115.7 (13.2)	114.7 (13.0)	116.6 (13.3)
	In Baseline Status	79.1 (17.6)	127.7 (27.8)	127.1 (27.9)	121.0 (27.2)	86.9 (23.5)
	In Rate of Change		.81 (.48)	.81 (.48)	.73 (.47)	.39 (.43)
Model Fit						
	Deviance	3026.8	3018.3	3017.4	3008.1	2982.8
	AIC	3032.8	3030.5	3033.4	3036.1	3030.8
	BIC	3044.6	3054.1	3064.9	3091.2	3125.15
	LR Test Chi-square (df)		8.3(3) *	1.1(2)	9.3(6)	25.3(10) **

Table 2.5: Multilevel Models for Health Problems Scale

The Demographics model does not change the picture substantially; the only significant coefficient is one indicating that older youth experience a greater decline in health problems than younger youth.

The deviance statistics indicate that neither the Demographics model or the Time+Site model is an improvement on its predecessor. The final model, including background conditions, does however have a number of significant terms and has more substantial evidence of an improvement in overall model fit. The indication that older youth have better improvement in health remains, and the lifetime substance use measure is significant for both baseline level and rate of change. The positive coefficient for baseline status and negative coefficient for slope mean that those with more severe substance abuse problems have more health problems at baseline, but also improve more than those with less severe substance use. This model also has a significant coefficient for the criminal behavior scale, indicating that those with more extensive criminal involvement are in better health, have fewer health problems, at baseline. This may be a case of reverse causation in that youth with serious health issues may be unlikely or unable to participate in criminal activity, or it could be a chance finding given the number of coefficients tested across these many models.

2.2.4 Maximum Change Score Analysis

As described in the Method section, I formed the maximum change score following the methods used by Boothroyd et al. (2004) using the same three outcome variables from the HLM models above. Figure 2.5 shows the distributions of the standardized change scores and their pairwise scatterplots. One appealing aspect of this approach is that the scores can be compared, since they are all in standard-deviation units. The change scores for substance use and emotional problems are

both heavier on the left side of zero, indicating improvement (i.e. a decrease in the problem measures), but the change scores for health are clustered slightly above zero and show less of a skew to the left compared to the other measures. In each plot there are a few outliers with high scores, indicating clients that got decidedly worse in the domain. The scatterplots show these outliers as well and an indication of a positive relationship between change in substance use and change in mental health, though much of that apparent relationship may be due to the more extreme points falling outside the cluster in the middle. The substance use and emotional problems scores are correlated, r=.34 ($p < .001$) but the others are not.

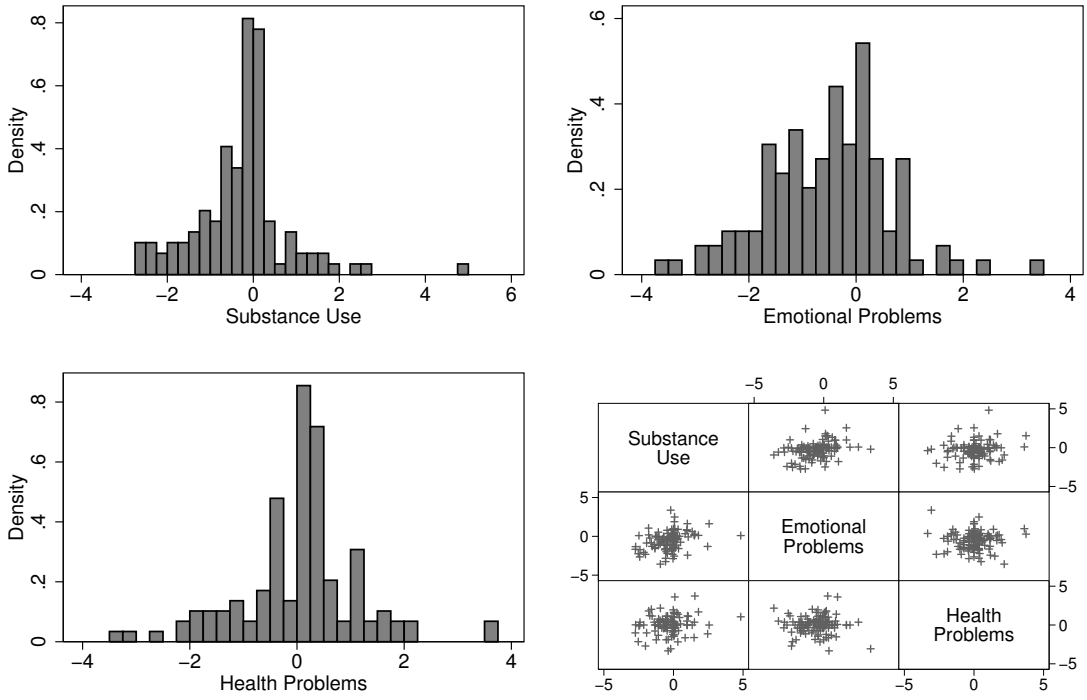


Figure 2.5: Distributions and Scatterplots of Standardized Change Scores

The core of the MCS approach is to take, within each client, the measure on which they improved the most, and form the outcome variable using those mixed values. Table 2.6 shows the breakdown of which measure was the maximally-improving one, overall and separately by site.

	Count(Pct)		
Measure	**Site A**	**Site B**	**Overall**
Substance Use Frequency	18 (32.1%)	21 (33.9%)	39 (33.0%)
Emotional Problems	25 (44.6%)	28 (45.2%)	53 (44.9%)
Health Problems	13 (23.2%)	13 (21.0%)	26 (22.0%)

Table 2.6: Maximally Changing Measures, Overall and by Site

Interestingly, for each measure there was a sizable number of clients for whom it was the maximally improving measure. Given that health problems were less of a central focus for the AAFT model, I anticipated a very small number of clients whom experienced their maximal change on this measure. Twenty-two percent of the clients changed the most on the health measure, 33% changed the most on the substance use measure, and 44% changed the most on the mental health measure. The other point to note about the figures in Table 2.6 is the nearly identical distribution of the maximally changing measure across the two programs. It appears that the blend of how each program helped its clients was the same across these different implementations of the AAFT model.

Figure 2.6 shows the distribution of the MCS overall and by site. Almost all clients experienced an improvement, (their change score is below zero) but for a small number their "best" score was in fact a small worsening in status (their "best" change score is above zero indicating an increase in a problem measure). The magnitude of the MCS is quite large, the mean and median are both near -1, meaning that the average client changed on *something* by about one standard deviation of the differences between clients at baseline. It is instructive to compare the distribution of the MCS to the standardized scores from which it was created, shown in Figure 2.5. The difference between the histograms in Figure 2.5 and the "Total" plot in this figure clearly shows the biasing effect of choosing the best score within each client. Though skewed towards the negative side, the histograms of the component scores are centered near zero, while the histogram of the MCS is

pulled strongly towards the negative, so that only a few clients are on the positive side. The distributions for the two sites show slightly different patterns. Site B has a fairly regular distribution centered around negative 1 while the distribution of scores in Site A is both wider, with a longer tail of clients experiencing large changes, and slightly bi-modal, perhaps indicating that at Site A there are sub-groups of clients that were differentially affected by the program.

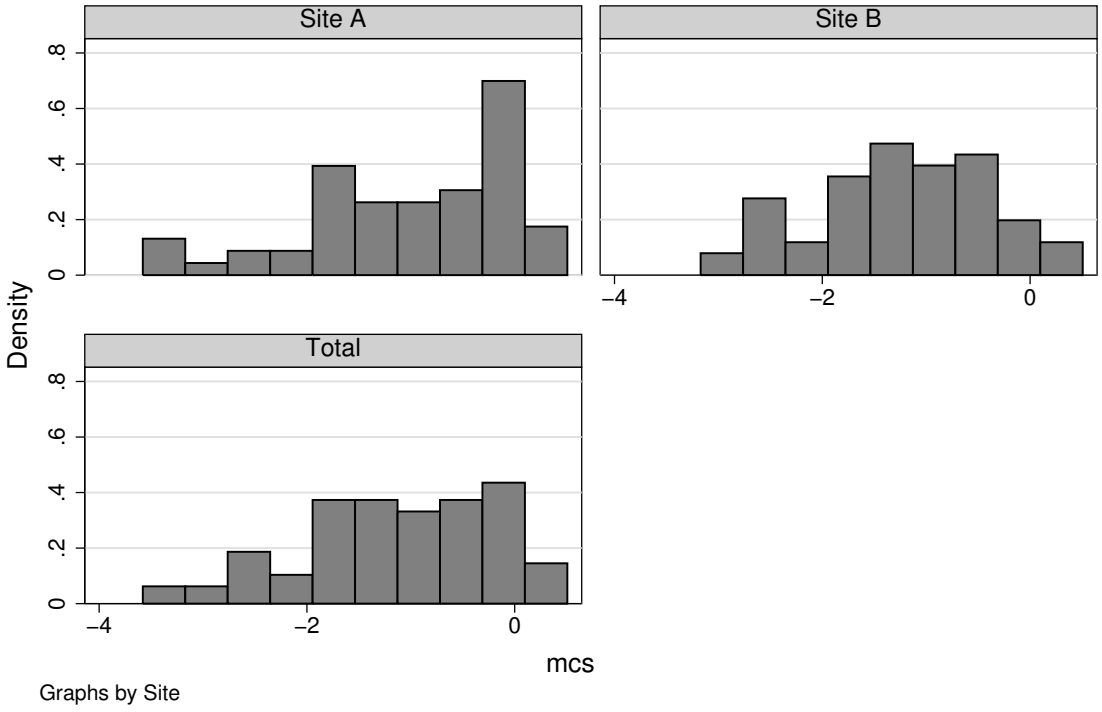


Figure 2.6: Distribution of Maximum Change Score Overall and By Site

Having created this maximum change score, we are able to compare the sites on it formally. Using the normal transformation for applying regular, distribution-based statistics, and the ranks of the scores for non-parametric tests, we can test whether the sites differed on the amount of change represented in their clients' MCS scores. Table 2.7 shows the results of the parametric (t-test) and non-parametric tests of the difference in change score between the two sites.

The tests indicate that clients in Site B may have experienced more change,

Test	**Site A**	**Site B**	**Statistic**	**p**
T-test	mean (SD): .26 (.03)	mean (SD): .18 (.02)	t(116) = 2.3	.02
Wilcoxon	rank sum: 3659.5	rank sum: 3361.5	z = 1.77	.08

Table 2.7: Statistical Tests for Differences in Maximum Change Score Between Sites

looking across these three measures at their "best" change, than clients in Site A. This simple approach however, leaves uncontrolled the potentially meaningful baseline differences in the populations served at the two programs.

To further examine whether clients experienced more change in one program than the other, I used the normal-transformed change score as the outcome in a series of simple linear models using the same predictors as in the HLM models above. Table 2.8 shows the results of fitting these models.

	Demographics	Dem. + Conditions
Measure	Coefficent (SE)	Coefficient (SE)
Intercept	.28 (.04) ***	.28 (.03) ***
Site	-.08 (.04) *	-.06 (.04)
Age 21 or older	-.11 (.04) **	-.10 (.04) *
Female	.01 (.04)	.02 (.04)
Nonwhite	-.00 (.04)	-.01 (.04)
Internal Mental Distress		-.006 (.002) *
Externalizing Problems		-.006 (.003) *
Victimization		.009 (.006)
Crime		.003 (.007)
Adjusted R^2	.07	.21
F-Test	F(4,113) = 3.1, p=.02	F(9,107) = 4.6, p<.001
LR test Chi-square		25.8 (5)***

Table 2.8: Models Predicting MCS

The first model includes site and the three demographic variables. This model conforms with the simple uncontrolled t-test above in finding that clients in Site B experienced more change than those in Site A, but this analysis controls for demographics. The coefficient for age is also significant showing that older youth experienced more change than those in the younger group. Once the background condition variables are added in the second model, the difference between sites

drops from .08 to .06 and becomes non-significant (p=.12) while both of the mental health measures are associated with greater change on the MCS. It appears that the difference between the sites could be due to higher mental health problems at baseline, which is associated with a greater *decrease* in the MCS. The mean mental health scores at baseline are higher in Site B than in Site A, supporting this interpretation.

2.2.5 Summary of Findings

The findings from the three hierarchical linear models and the model for the maximum change scores are summarized in bullet form below.

In the models for substance use frequency the initial differences by site that were found disappeared once background conditions are controlled for:

- Clients in Site B have higher levels of baseline use than those in Site A.
- Clients with higher mental health problem scores at baseline have higher use at baseline.
- Clients with higher levels of victimization have *lower* use at baseline
- When background factors are entered into the model, differences between sites disappear
- Clients with higher internal mental distress at baseline show greater declines in substance use
- Clients with greater victimization at baseline show smaller declines in substance use

In the model for the emotional problems outcome the only significant relationships that were found were between the predictors and baseline levels of the outcome. The specific findings were:

- Clients with higher baseline levels of substance use have higher baseline levels of emotional problems
- Clients with higher baseline levels of victimization have higher baseline levels of emotional problems
- Neither site, demographics, nor background factors predicted rate of change

Health problems had the anomalous relationship to criminal justice involvement, and the more easily interpretable relationship to substance use problems:

- Clients with higher levels of criminal involvement have *better* health at baseline
- Older clients have greater improvements in health
- Clients with more severe substance use problems at baseline start in worse health but improve more

The maximum change score variable ended up being a composite of all three of the separate outcomes, substantial proportions of clients had each measure as their maximally changing outcome. As with the substance use model, in the MCS model initial differences between the sites disappeared when baseline differences in the populations at baseline were controlled.

- Emotional problems was the variable on which the largest share of clients changed the most (45%), followed by substance use (33%), and health (22%).
- The distributions of the most-changing variable were similar in the two programs
- Simple statistics indicate that clients in Site B experienced more change
- When modeled, differences in site disappear
- Age is associated with greater improvements on the MCS

- Higher baseline levels of mental health problems are associated with greater improvements on the MCS

CHAPTER 3

STUDY B: WITHIN-PROGRAM EXAMINATION OF DIFFERENTIAL EFFECTIVENESS

Study B focuses on within-program differential effectiveness and draws upon data from the evaluation of Project Genesis, a project in Boston designed to help homeless women improve their well-being and gain stability.

3.1 Methods

3.1.1 Program and Data

In 2006 a Boston area non-profit agency serving the homeless received a five-year grant under the Substance Abuse and Mental Health Service Administration's Treatment for Homeless program to implement Genesis. The core service component of Genesis was intensive, flexible case management provided by a team of "Care Coordinators" who conducted outreach, worked to develop trusting relationships with homeless women, and then brokered a wide range of services including help accessing housing, help obtaining benefits, counseling, substance abuse and mental health services, motivational interviewing, and flexible, instrumental supports such as food and bus passes. The program also included a focus on trauma services and the Care Coordinators received training and ongoing support from specialized clinicians for counseling women who have experienced trauma.

As a condition of funding, Genesis was required to conduct client interviews at baseline and six months following enrollment. These interviews were conducted by evaluation staff from a separate agency using both a government-mandated short assessment tool and a much broader set of standardized research instruments. As part of the evaluation, the Care Coordinators recorded their service contacts with

Genesis clients, uploading records monthly that included the date of each contact, the client it was with, the particular topic(s) of the contact, and the specific services delivered.

3.1.2 Measures

Project Genesis was designed to potentially impact multiple areas of homeless women's lives and the evaluation interviews accordingly assessed multiple domains. Across both the government-mandated short assessment, and the longer assessment developed for the evaluation, the following major domains were measured on participants at enrollment and six months following enrollment.

- Residential stability
- Mental health status
- Substance use
- Traumatic experience and trauma symptom severity
- Employment, income, and benefits
- Legal issues and criminal justice involvement
- Self esteem
- Health status
- Social support

The first four of these were the central domains the program was hoping to impact and I chose the first two, residential stability and mental health status, for outcome modeling. Residential stability was included because it is of course central to overall stability for homeless women and it may facilitate other changes in women's lives. While Project Genesis did not control housing, i.e. had no housing resources available for its clients, there were improvements for women in this area.

The specific item used as the outcome is the response to a question that asked women "In the past 30 days, where have you been living most of the time?". The responses were dichotomized into "shelter" or "street" on the one hand, and all other locations on the other. The shelter/street response was coded as a 1 and other responses as 0 so this outcome indicates having had one's primary living location over the past 30 days be shelters and/or the streets. It is thus a measure of "literal homelessness", among a group of women who were all homeless, but lived in a variety of (frequently changing) situations.

The second domain, mental health, was chosen because it is also of critical importance, is potentially amenable to the sorts of treatment being offered by Genesis, and it has a strong continuous measure. For these analyses I wished to have both a continuous and a categorical outcome to compare across the two methods. The specific outcome is severity of symptoms from mental illness, as measured by the Brief Symptom Inventory (Derogatis & Melisaratos, 1983). This measure asks clients how much during the past week they've been distressed by 53 different symptoms and has been shown to have good reliability (Boulet & Boss, 1991) and is sensitive to change over short time periods.

Besides these outcomes, the modeling analyses include demographics, service use measures, and measures of background conditions as predictors. The demographics are age, race, Hispanic/Latino origin, and education level. The service use measures are the Care Coordinator the woman worked with and the number of service contacts the woman had across her involvement in the project. There are four measures of the presence/severity of background conditions indexing the domains of mental health, homelessness, substance use, and trauma. The background measure for mental health is having ever (over the lifetime) been admitted to an inpatient facility for mental/emotional issues. The homelessness background mea-

sure is duration of homelessness as an adult, with values of less than two years, two to five years, and five years or more. The substance use measure is the recency of drug/alcohol use with categories of less than one month ago, one to three months, and four or more months. The trauma measure is a count of types of traumatic events experienced. There are twelve events assessed, covering disasters, illnesses, witnessing assaults, death of important others, and several forms of physical and sexual violence (see Table 3.3).

3.1.3 Pre/Post Linear Models

A typical way to look for subgroup differences is to model the outcome of interest, measured at a post-treatment time point, as a function of its baseline value, and the covariates that one suspects might impact the outcome. Under this approach the baseline value of the outcome variable is included to control for the client's starting place. As described in the Introduction, there has been a long debate in the statistical literature over the best approach to such pre/post analysis with covariates, and modeling the post-value as a function of pre-value plus covariates is recommended. This allows the analyst to determine whether covariates are predictive of final status while controlling for baseline status.

In these models I model follow-up status as a function of the corresponding baseline status and two sets of covariates: service use variables and background conditions. These models are thus examining whether clients ended up in different places, net of their starting place, due to service use or to background conditions. As described above, there are four measures of background condition severity that index the domains of mental health, homelessness, substance use, and trauma. The models include these background measures, but each model excludes the one that is for the same domain as the outcome, as this would likely be collinear and

redundant. So for example, the first model described below predicting follow-up mental illness symptoms includes the baseline value of the mental illness symptoms measure plus the lifetime/background measures for homelessness, substance use, and trauma, but not the lifetime/background measure for mental health.

In fitting these models, I have attempted to balance inclusion with parsimony. Statisticians recommend different strategies for building models. Some (e.g. Hosmer & Lemeshow, 1989; Crawley, 2002) emphasize parsimony in modeling and others (e.g. Gelman & Hill, 2007) favor broader inclusion of possibly relevant terms and interactions. Because we are interested in examining patterns of differential effectiveness rather than in developing models of the outcomes *per se*, it seems prudent to generally favor inclusion over parsimony and adopt a relatively broader strategy. Accordingly, in each of the models here I include a designated set of predictors, rather than following a more parsimonious strategy of first screening candidate predictor variables according to their bivariate relationships to the outcome. But, because we are trying to answer a question that inherently has to do with the question "what variables matter?", and to come to a definitive understanding of that question, I attempted to enhance parsimony by eliminating non-significant terms from models, rather than leaving them in. In order to be broadly inclusive in this effort, I used a generous significance level, .10, to make sure important predictors are kept.

3.1.4 Trees and Forests

The Introduction chapter describes the tree-fitting methodology conceptually. This section discusses the features of the specific algorithm used. As mentioned in the Introduction, I use here the *conditional inference* algorithm of Hothorn et al. (2006). This algorithm is distinguished from earlier Classification and Regression

Tree (CART) methods by using statistical criteria to determine the "importance" of each potential split. Original CART methods have no notion of statistical significance. With standard CART methods, as the algorithm searches for variables to split on, and the splitting values to use, it simply finds the variable and value of that variable that best distinguishes cases on the dependent variable. This procedure is purely mathematical, finding the values that lead to a maximal difference between the resulting groups in terms of the dependent variable.

The incorporation of statistical criteria into the splitting rules improves over older methods in two ways. First, by computing at each step a test statistic that indexes the association between each candidate variable and the outcome, and comparing only these test statistics, this method overcomes a problem with earlier methods in that they tended to choose variables with many possible splits (i.e they preferred continuous measures over categorical ones). Second, by using statistical criteria the procedure is able to apply standard multiple test procedures and determine whether across the set of possible predictors no significant association between them and the dependent measure exists, and the splitting procedure should stop. Standard CART methods continue until all subjects are classified into homogeneous groups, or are put into a group by themselves, and no more splitting can be done. The analyst then has to "prune" the tree to avoid over-fitting. With the conditional inference method, however, each potential split is evaluated with overall statistical criteria and the method stops when the null hypothesis of no association between the predictors and outcome cannot be rejected.

As described in the Introduction, the random forest analysis proceeds by conducting many individual tree analyses and averaging over the results. To increase the variability of the trees produced, each tree is run on a subset of the overall set of predictor variables. By making the trees more diverse and then averaging

over them, the method increases the chances that the findings are not due to particularities of the data in question. For the forest approach there are three key parameters that need to be specified: the number of trees to run, the number of variables to include in each tree, and the minimum number of cases to be allowed to be considered a group. For each forest analysis I ran 1000 trees. These results took only a few seconds to produce and seemed stable across multiple runs. One unusual feature of ensemble methods is that one can obtain different answers each time the analysis is run, since random selection of participants in sub-samples and random selection of variables is used. It is standard practice to set the number of variables to include in each analysis to approximately the square root of the total number of variables, which I followed here, setting it to four. Because it might be of value to find small, but statistically important subgroups of women in the Genesis data, I set the minimum number of cases per group to five.

The forest method cannot produce easily interpretable output as is possible with single tree methods, since different trees in the forest have different subsets of variables, and variables appear in all different positions in the different trees, so no overall graphic or other representation can easily be made summarizing the results. The *conditional inference* forest method used here (Hothorn et al., 2006) can however produce measures of overall variable importance across the trees of the forest. The method accomplishes this by an ingenious technique, permutation tests. Each predictor variable is randomly re-arranged within itself, the cases are randomly re-ordered, or permuted, and each tree is re-fit using this altered version of the variable. If the predictor has an association with the outcome, then randomly rearranging it breaks this association and the tree model fit with the rearranged version will have lower predictive accuracy than the version fit with the regular variable. The average decrease in accuracy across the trees seen by rearranging

the predictor is a measure of its importance. Because this measure is dependent on the specifics of the analysis it cannot be compared across studies, but within the forest analysis it provides a relative measure of variable importance.

3.1.5 Provider Feedback

To gage provider reactions to the pre/post models and the tree methods, I conducted a focus group with four of the staff from Project Genesis. Overall six core staff worked on the project over the five years so this group should be fairly representative of the staff's views. In the group I presented the findings from the two analyses in a simple format (see Appendix A) and asked the staff to discus the findings. Before presenting the findings I asked the providers to reflect on their own intuitions regarding the program's differential effectiveness. After the results were presented to the providers, I asked them to discuss four aspects of the results:

- Understandability: How easy is it to understand the results provided by the method?
- Scope: Are the results at an appropriate level and neither too detailed nor too global?
- Validity: Are the results believable, do they have "face validity"?
- Utility: Do the results seem potentially useful in terms of practical applications and clinical practice?

The specific text of the questions is provided in the Focus Group Protocol (Appendix B). At the conclusion of the focus group I asked the participants to complete a simple single-page questionnaire (Appendix C) quantifying their answers to the above questions for each of the two methods.

3.2 Results

The question to be examined for this second study concerns differential effectiveness, whether, when looking within a program, we can discern different sub-groups of clients who show more, or less, change over time than others. The first step of any sub-group analysis must be to determine whether the client population varies to an important extent on these causally prior factors and on the outcomes of interest. The following sections contain descriptive analyses of the background and outcome variables, preliminary to the subsequent sections that contain the results from the different analytic methods.

3.2.1 Profile of Genesis Participants' Background and Service Use

Although by its definition the target population for Project Genesis, single women experiencing homelessness in Boston, was relatively narrow, there was a good deal of variability in the women's background characteristics. Table 3.1 shows demographic characteristics of the program participants.

The women in Genesis ranged from 19 to 62 years old with a mean of 41, much older than is typically seen with women in homeless families, who are often in their 20's. This age range is more typical of that seen with single adult homeless men, so the women in Genesis may resemble the male single adult homeless population more than they resemble the typical woman in a family that is homeless. The sample was racially diverse with two large groups: white (41%) and African American 32%), with 16% Hispanic (of any race). A problem with data collection during the study's initial phases resulted in race information not being collected for a subsample of participants. Almost one third of the women (27%) had some

Measure	Mean(SD)[range] or Count(Pct)
Age	40.9 (10.2) [19-62]
Race	
White	60 (41%)
Black	47 (32%)
Other/Multiple	14 (10%)
Refused/Not Ans.	26 (18%)
Hispanic/Latino	23 (16%)
Marital Status	
Never married	83 (59%)
Divorced/widowed/separated	46 (33%)
Married	11 (8%)
Education	
Some HS or less	34 (23%)
High school	73 (50%)
Some college	27 (18%)
College or more	13 (9%)
Monthly income	436 (373) [0-1800]
Income sources	
Wages	10 (7%)
Public Assistance	46 (31%)
Retirement	1 (1%))
Disability programs	65 (44%))
Family and friends	20 (14%))
Other	4 (3%))
Children	111 (76%)
Lost custody (among those w/ children)	29 (27%)

Table 3.1: Demographic Characteristics of Genesis Participants

higher education, but despite this the group is exceedingly impoverished, with a mean monthly income of $436. The distribution of income in the sample is strongly bimodal with one group near 0 income and another receiving around $600/month. This second group is made up of women receiving disability benefits. Most of the women were mothers, though of these 27% had lost custody of their children. While the adult homeless population is often conceptualized as single adults vs. families, these figures underscore that "single adults" often have familial connections.

Tables 3.2 and 3.3 profile participants on key background issues and traumatic

experiences. The first four measures in table 3.2 concern residential instability. Women in Genesis generally had severe homeless histories; almost two thirds report homeless durations as an adult of 2 years or more. Many of the women had their first homeless spell relatively later in life (mean age = 31.5), again indicating that these women differ substantially from homeless women in families. Although homeless histories for the group as a whole are quite severe, there is still variability on these measures – about 10% of the women have only short spells in homelessness, 6 months or less, and the range in the number of times women have experienced homelessness is quite large, from 1 to 30. Sizable subgroups of women spent time in foster care or group homes as children and approximately half have received treatment for acute mental illness in their lifetimes. The age of first using alcohol to intoxication or illegal drugs, considered a good marker for likely substance abuse problems later in life, is quite young, on average 15 years.

Measure	**Mean(SD)[range] or Count(Pct)**
Times homeless as an adult	3.8 (4.3) [1-30]
Age first homeless	31.5 (11.9) [11-57]
Lifetime duration homeless	
Less than 1 month	3 (2.4%)
1 - 6 months	9 (7.3%)
6 months - 1 year	15 (12.2%)
1 - 2 years	17 (13.8%)
2 - 5 years	39 (31.7%)
5 years or more	40 (32.5%)
Doubled up as an adult	121 (82.9%)
Foster care as child	29 (20.0%)
Group home as child	23 (16.1%)
ER tx. for psychological issue	77 (53.8%)
Inpatient tx. for psychological issue	74 (52.9%)
Age of first mental health tx.	24.3 (12.7) [3-54]
Age first getting drunk/using drugs	15.0 (5.8) [0-53]

Table 3.2: Background Conditions of Genesis Participants

Any profiling of homeless women would be incomplete without an examination

of the pervasiveness of violence and other traumatic experiences in these womens' lives. Table 3.3 lists the prevalence of twelve traumatic experiences among the Genesis population.

Traumatic Experience	**Count (Pct.)**
Serious disaster	49 (33.8%)
Life-threatening accident/illness	51 (35.2%)
Family/partner/friend died traumatically	62 (44.0%)
Death of a child	24 (18.1%)
Present during physical/sexual assault	67 (46.5%)
Physical violence from family member/known	74 (51.8%)
Physical violence from partner	105 (73.4%)
Physical violence from stranger	75 (52.5%)
Strip searched/restrained	78 (54.6%)
Sex in exchange for money, drugs, shelter	69 (48.9%)
Sexual assault by family member/known	77 (54.2%)
Sexual assault by stranger	55 (39.0%)

Table 3.3: Traumatic Experiences of Genesis Participants

As the table makes clear, traumatic experiences are normative among chronically homeless women. Tragically, 24 women reported that a child of theirs had died. The rates of physical and sexual violence are especially high: over 70% have experienced intimate partner violence and 88% have suffered from at least one of the five types of physical or sexual violence measured.

Although all of the women in the sample technically participated in the program, with a flexible, unstructured, participant-driven program such as Genesis, what it means to "participate" could potentially vary widely from woman to woman. Analysis of the service contacts data reveals that women did in fact vary greatly in their degree of engagement with the program. First, because the project lasted five years and had no defined course for participants to follow, and no definitive exit criteria, the range of time that women spent in the program (defined as the number of days between their first and last service contacts) varied greatly. Besides this variability, women engaged in the program at differing intensities.

Figure 3.1 shows the distributions of time in the program and service contacts per month.

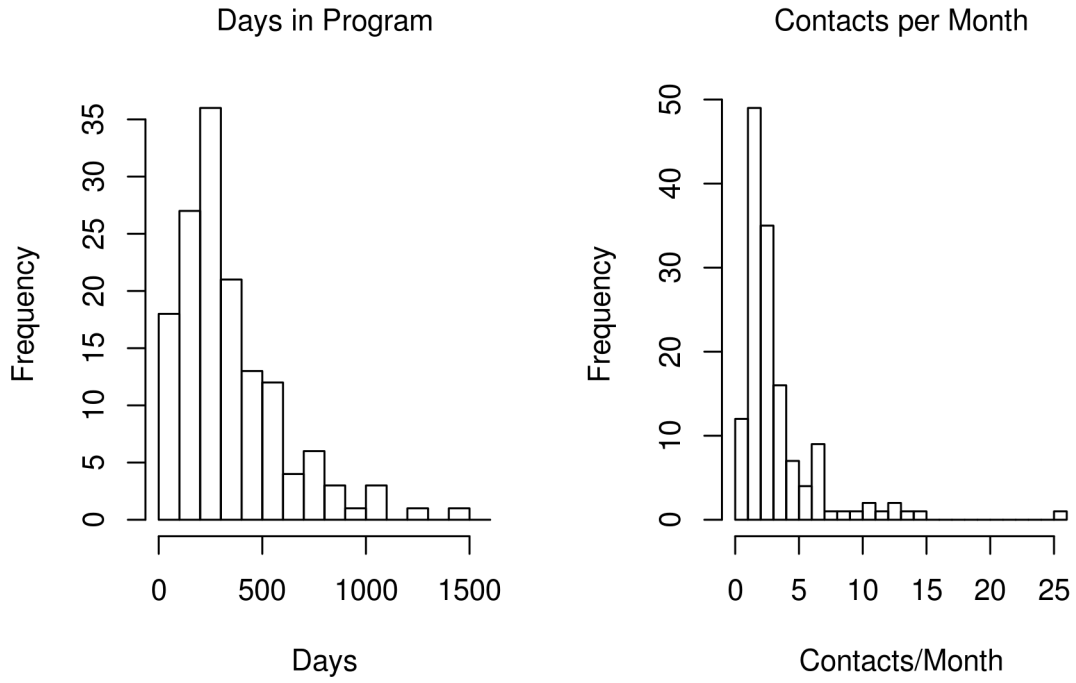


Figure 3.1: Service Engagement Among Genesis Participants

The median time of engagement in the program is 282 days, but as the histogram makes clear, there was wide variability in the lengths of engagement. Two thirds of the women were engaged with the program for less than a year, but a small portion engaged over a period of several years. Because the length of engagement is constrained by when women enrolled and therefore how long they had an opportunity to remain in the program before it ended in 2011, the number of service contacts per month provides a potentially more useful measure of service engagement. While this measure also shows some variability, we see a narrower distribution, with the women relatively tightly clustered around 2-3 contacts per month (the median is 2.3), though again there is a significant minority of women who maintained more contact with the program; About a quarter of the women had

more than one contact per week. The higher level of uniformity may be because of the Care Coordinators' efforts to maintain ongoing contact with women, and therefore this measure is less subject to the individual situations and conditions of the participants.

One potentially important factor in participants' engagement is the specific front-line "Care Coordinator" with whom they worked. As described above, Project Genesis employed a total of four front line service providers, termed Care Coordinators, over the five years of the project in three funded slots, representing an astonishingly low level of turnover (i.e. only one slot turned over over the five years of the project. The four Care Coordinators did not see equal numbers of clients. Table 3.4 shows the number of clients for whom each of the four staff was the primary Coordinator.

Care Coordinator	Clients - Count (Pct.)
A	16 (10.9%)
B	24 (16.3%)
C	99 (67.3%)
D	8 (5.4%)

Table 3.4: Genesis Care Coordinators' Caseloads

Care Coordinator B left the project in its fourth year and was replaced by D, and A and C stayed with the project all five years. Coordinator C, who primarily recruited clients from shelters, had by far the largest caseload. Coordinator A recruited women directly from the streets, and often worked for women over a long period of time to gain their trust before they were comfortable joining the program. Coordinators B and D were charged with recruiting women from other homeless agencies, and this proved difficult for the program as a whole. Because of the widely varying caseloads, we see varying levels of engagement on the part of participants with the different Coordinators.

Figure 3.2 shows the number of service contacts per month separately by Care

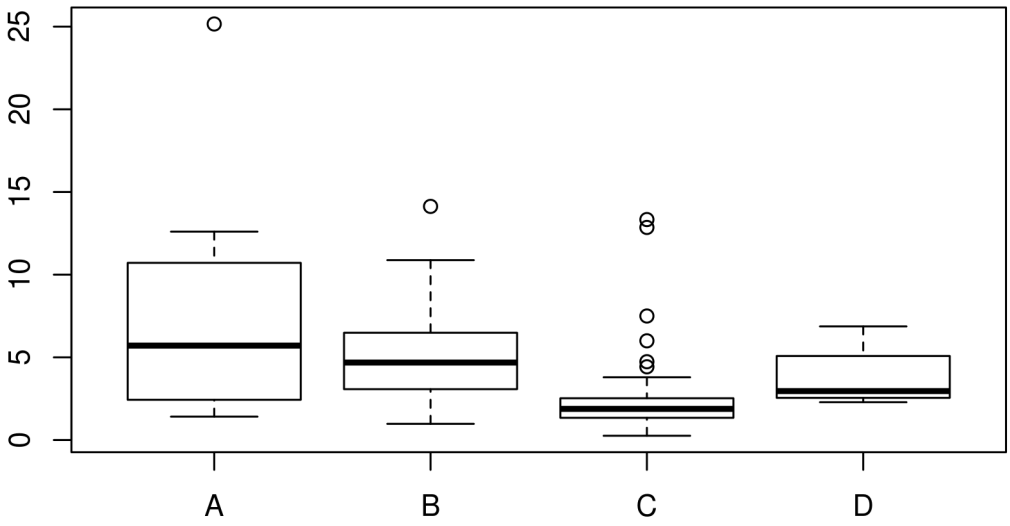


Figure 3.2: Distribution of Contacts per Month By Care Coordinator

Coordinator. Coordinator C's clients had fewer contacts and were much more tightly clustered near roughly two per month than the other Coordinators'. So for women engaging with different Coordinators, it is possible that the experience of Project Genesis was very different.

3.2.2 Change on Outcome Measures for Genesis Participants

In general, the Genesis evaluation found the program to be effective. Genesis women experienced statistically significant and clinically meaningful improvements on multiple measures between the baseline and follow-up assessments across all four of the primary domains of interest: residential stability, substance use, mental health, and trauma, as well as others. At follow-up women were more likely to

be living in their own home, less likely to be literally homeless, less likely to have used alcohol or drugs, and less likely to have experienced negative consequences of drug/alcohol use. They also reported improved psychological and trauma symptoms, higher income, and increased employment. There was a non-significant trend towards improvement in health (p=.065) but a decline in self-esteem. While the Genesis evaluation was limited by its lack of a comparison group and the short time span between the baseline and follow-up assesssments, and we cannot strongly infer causality between the program and the changes in women's lives, it does appear that the participants at follow-up were in substantially improved conditions compared to their status six months previously.

As described under Measures above, this study focuses on the most important measures from two of the most important domains assesssed: residential stability and mental health. On the Brief Symptom Inventory, that measures severity of mental illness symptoms experienced in the past week, the mean score dropped from 65.1 to 61.1, on the scale that has a possible range of 33 to 80. This change is statistically significant ($t(142) = 4.42$, $p < .001$), and large enough to be clinically meaningful. The panels in Figure 3.3 show the distribution of scores at baseline and at follow-up (top two panels), and the distribution of the change scores (bottom left). The fourth panel of the figure (bottom right) shows the individual change scores for the 147 women, sorted from positive (largest increase in symptom severity), to negative (largest decreases in symptom severity). Plotting the change scores in this fashion could be useful because it gives a visual indicator (the shaded areas) of how much change was experienced by how many clients in each direction. Women in Genesis had a broad range of mental health symptom severity at baseline, with most scoring in the upper end of the range. Clinically, scores above 63 are considered indicators of significant impairment. The finding of an overall

decrease in symptoms is evident in the shifting between the top histograms from right to left, and in the modal change score being below 0. If one focuses on the shaded areas in the final panel one can in effect get a sense of the "overall amount of change" (defined as how much change experienced by how many clients) and see that, as the statistics support, the change in the negative direction (i.e. a decrease in symptoms) is substantially larger than the increase in symptoms, both in width (how many women experienced it) and in average height (how large the change per person was).

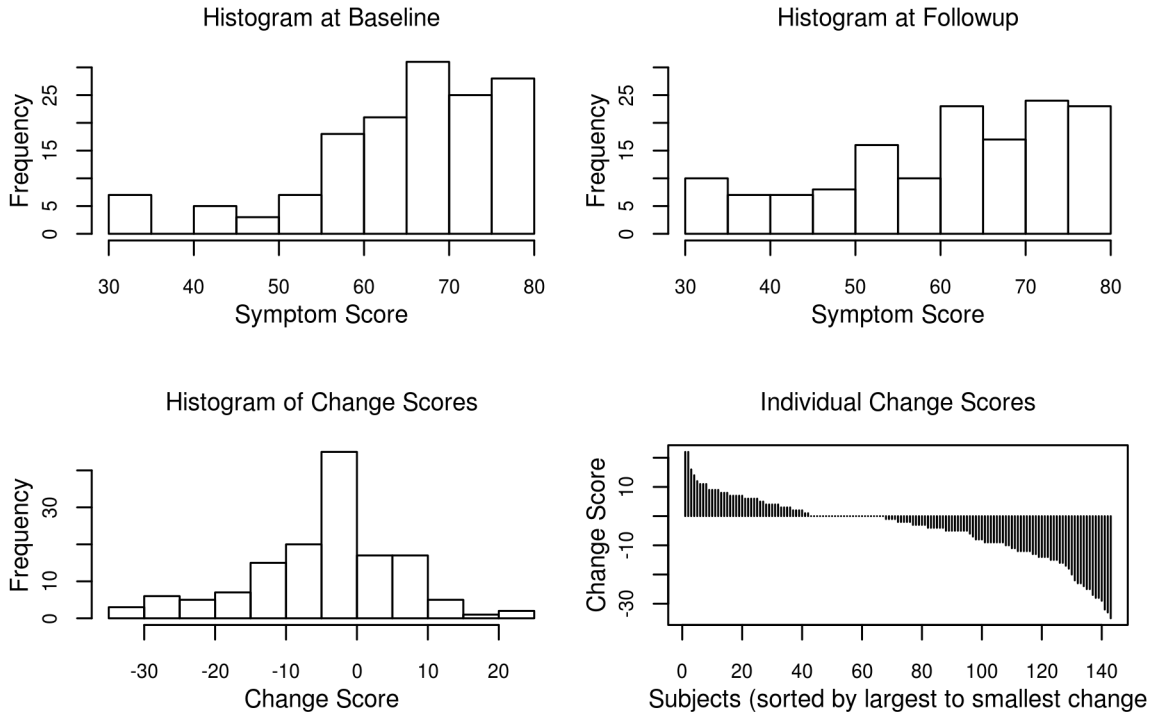


Figure 3.3: Change In Genesis Participants' Mental Illness Symptoms

The second measure, having a predominant living location of shelter or streets, also showed significant change from baseline to follow-up. At baseline, 54.2% were living in shelter or on the streets, while at follow-up this had declined to 41.5.02). The majority of women who were not literally homeless at baseline were either doubled-up (15.8%) or in some type of institution (14.4%). Table 3.5 shows the

frequency of the four possibilities from crossing the baseline and follow-up housing status measures.

Housing Status	Count (Pct.)
Homeless at neither timepoint	47 (33.1%)
Homeless at both timepoints	41 (28.9%)
Became homeless	18 (12.7%)
Stopped being homeless	36 (25.4%)

Table 3.5: Change in Genesis Participants' Housing Status

Interestingly, there was no correlation between the two outcomes. The point-biserial correlation between literal homelessness and mental health symptoms at follow-up is r(140)=.02.

3.2.3 Pre/Post Linear Models

This section contains results of fitting the linear pre/post models.

Mental Illness Symptom Severity at Follow-up

The first model is for the Brief Symptom Inventory's Global Severity Index, which assesses severity of symptoms from mental illness. After fitting the full model, removing non-significant main effects, and testing for two-way interactions, I arrived at the final model predicting mental illness symptoms that is reported in Table 3.6.

As expected, baseline score is strongly related to follow-up score, each one point rise in baseline score is associated with a 0.72 rise in follow-up score. Being older and having a more recent drug/alcohol use problem are also marginally related to follow-up symptom severity in the expected direction, indicating that these groups experienced less positive change in the mental health area. Having experienced homelessness for 2-5 years was predictive of higher mental illness symptoms com-

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	58.64	2.00	29.3	0.000
Baseline symptom severity	0.72	0.08	8.7	0.000
Age	0.19	0.09	2.0	0.044
Lifetime hmls. 2-5 yrs	4.91	2.37	2.1	0.041
Lifetime hmls. 5 + yrs	-1.00	2.41	-0.4	0.679
Last use 1-3 mo.	-1.70	2.37	-0.7	0.475
Last use < 1 mo.	4.37	2.23	2.0	0.052
Traumatic evt. count	0.56	0.34	1.7	0.100

Table 3.6: Linear Model of Mental Illness Symptom Severity

pared to the reference category of having been homeless for less than two years. Interestingly, this was not the case for those with extensive homeless histories of five years or more. This model provides suggestive, but not strong evidence, that women who were older, in the mid-range of homeless experience, and recent users were perhaps less likely to have positive mental health outcomes than others.

Homelessness at Follow-up

The second outcome measure is living in shelter or on the streets at follow-up. As this is a dichotomous outcome, it is fit with a logistic regression model, which is presented in Table 3.7.

	Odds Ratio	Std. Error	z value	Pr(>|z|)
(Intercept)	0.04	0.49	-2.8	0.00
Shelter/street @ baseline	3.17	1.32	2.8	0.01
Age	1.04	0.02	1.9	0.05
Care Coordinator B	0.08	0.08	-2.4	0.01
Care Coordinator C	1.90	1.33	0.9	0.36
Care Coordinator D	0.80	0.90	-0.2	0.85
Number svc. contacts	1.01	0.01	2.2	0.03

Table 3.7: Logistic Model of Homelessness at Follow-up

As in the model for the mental health measure, the baseline value is a strong predictor of the outcome; here women who were homeless at baseline were three

times more likely to be homeless at follow-up than women who were not. Older women were more likely to be homeless at follow-up. The odds ratio of 1.04 per year of age implies that for every 10 years of age, the odds of being homeless go up by about 50%. The significant coefficient for one of the Care Coordinators is difficult to interpret as it is, since it is a comparison of one Care Coordinator (B) to another (A, which is the reference category). If we test the overall strength of the effect of Care Coordinator, it is statistically significant ($\chi^2(3) = 9.8$, p=.02. Pairwise tests of the means indicate that the clients of Care Coordinator B were less likely to be homeless than clients of the other Coordinators. The Care Coordinators recruited from different institutions and settings, and this difference is almost certainly due to that contextual factor. Care Coordinator B recruited primarily from a long-term medical care facility, so it is likely that many women with whom she worked were still in the facility at the time of the follow-up interview. Interestingly, having higher numbers of service contacts was associated with *increased* odds of being homeless at follow-up, perhaps because women with access to housing through other means were less likely to stay engaged with the program.

3.2.4 Tree and Forest Analyses

To explore the tree-based methods, I fit tree models to each of the outcomes from above using the same set of independent variables as used in the linear models. The results from the tree analyses are usually presented graphically, in a tree diagram format that can show quite intuitively the results of the model.

Mental Illness Symptom Severity at Follow-up

The results for fitting a tree using the bias-free methods of Hothorn et al. (2006) are shown in Figure 3.4. The figure is read from the top down. Each oval represents a

node where the algorithm has split a group into two groups, based on the particular value of the variable indicated. Within each oval is the p-value indicating the statistical significance of the improvement in fit obtained by making the split. The rectangles at the bottom of the tree represent the final partition of the clients into groups. Within each rectangle a box plot shows the distribution of the dependent variable, here follow-up mental illness symptom severity, within the group.

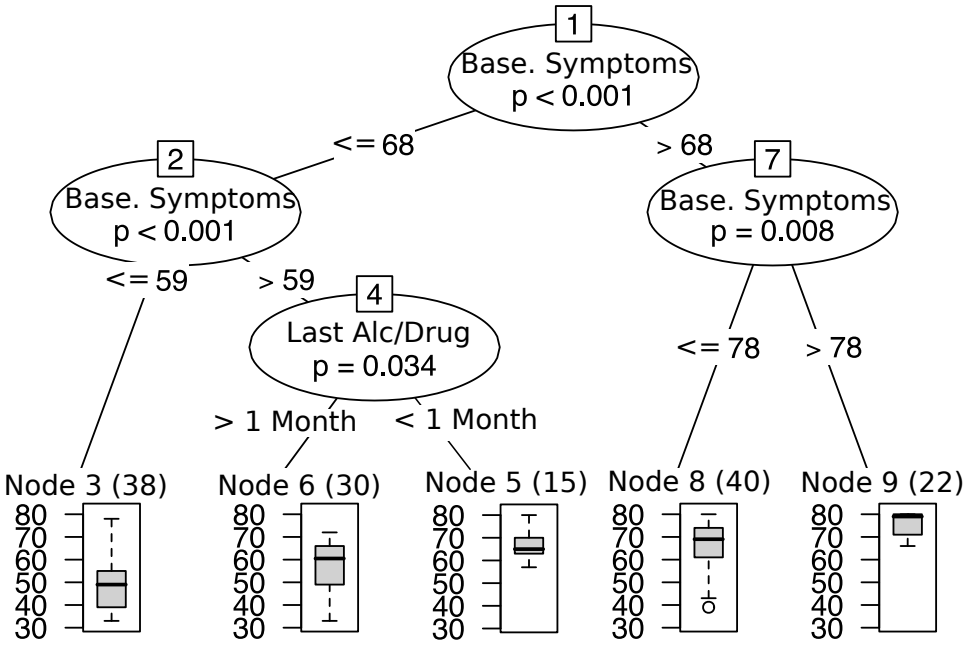


Figure 3.4: CART Analysis Predicting Mental Illness Symptom Severity at Follow-up

In the first node all of the clients are split on the basis of their baseline symptom severity scores into two groups, those with scores less than or equal to 68, and those with scores over 68. Interestingly, this split is not far from the clinical cutoff representing "severe" symptoms that is often used with the GSI, which is 64. Those with lower scores are then further sub-divided into those with scores less than or equal to 59 vs. those with scores greater than 59. On the right side of the tree, the

initial group with high baseline scores is also further split into two groups based on baseline score at the cutoff of 78. All three of these splits represent the expected effect of baseline severity scores in shaping the follow-up scores. Just as baseline score was naturally highly predictive of follow-up score in the linear models above, the first, most important splits the tree algorithm finds are on the basis of the baseline value. The middle group in the tree is those with mid-range baseline severity scores, from 59 to 77. In this mid-range group, recency of substance use becomes important. Those with use less than 1 month ago (basically currently using) are split from those whose use is more in the past, and this former group has statistically significantly higher mental illness scores at follow-up. This makes some intuitive sense, as women with either very positive or very negative mental health have their follow-up scores strongly determined whereas for women in the mid-range the scores are less determined and there is more "room" for other factors to intercede. So this model predicts that women with low baseline scores will have low follow-up scores, women with high baseline scores will have high follow-up scores, but for women in the mid-range at baseline, those who are currently using will have worse mental health at follow-up than those whose use is further in the past.

The random forest analysis produced a slightly different set of findings, though it is difficult to interpret because of the lack of a single fitted solution, as discussed above. The forest analysis produces a relative importance measure for the variables (see Methods). While this measure cannot be compared across forest analyses, it provides a basic ranking of how important each variable was across the multiple (in this case 1000) trees fitted. As expected, the baseline severity score is by far the most important predictor, with a permutation importance score of 97.5. The only other measures with values appreciably greater than zero are, in order, the

count of traumatic events experienced (score=15.9), the recency of drug/alcohol use (score = 6.2), and the duration of homelessness (score = 5.9). This analysis did not reproduce the single tree finding of recency of use being the most important predictor after baseline severity. Instead, it finds that the count of trauma experiences is more important, and by a significant margin. This may be due to a "masking" effect, that the random forest analysis is designed to overcome (see Discussion).

Homelessness at Follow-up

The tree model for the homelessness outcome is presented in Figure 3.5. The format of the diagram is the same as with the mental illness measure above, but in this case the outcome id dichotomous instead of continuous. Therefore, the rectangular boxes representing the terminal nodes, the final grouping of clients, do not contain box plots of the outcome but instead a simple bar-graph representation of the outcome, in this case the proportion of women in the group who were homeless at follow-up.

The algorithm has found the same basic structure as with the mental health measure, and the first split is naturally on the baseline value. Women who were living in a shelter or on the streets at baseline are split from those who were not. Within those who were living in a more stable situation at baseline, the number of types of traumatic events experienced becomes important. Those with very extensive trauma histories, who experienced eight or more of the 12 possible types of traumatic events on the measure, are much more likely to be homeless at follow-up than those who were also stably housed at baseline but had less trauma history. This latter group, women who were not literally homeless at baseline and who have (relatively) lower trauma histories are relatively unlikely to be homeless at follow-up.

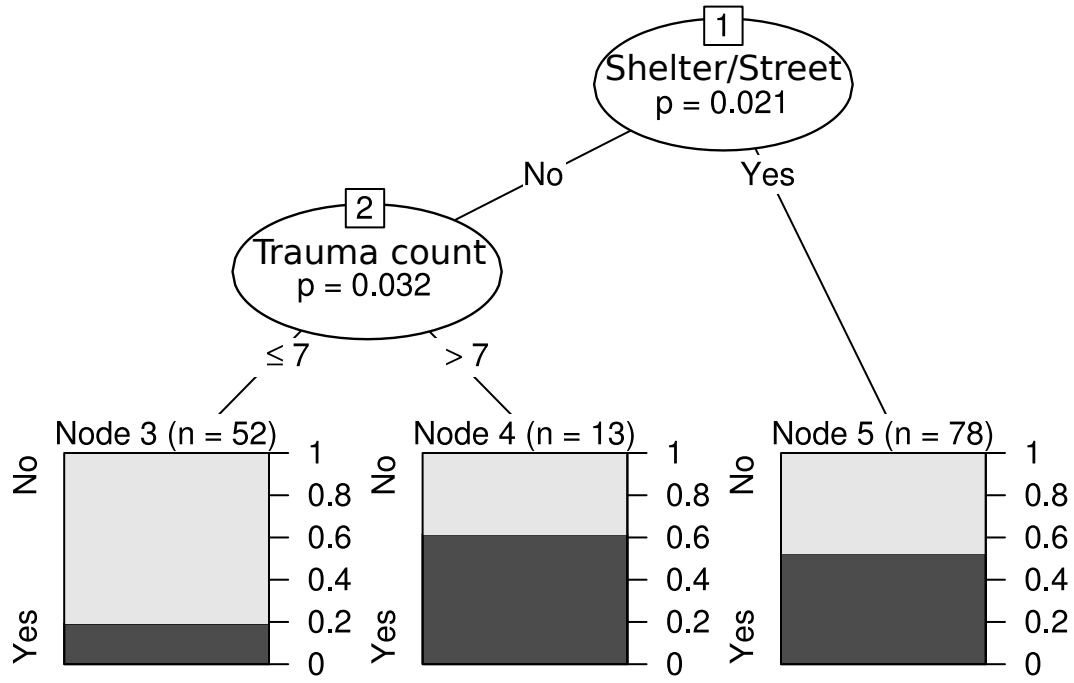


Figure 3.5: CART Analysis Predicting Homelessness at Follow-up

The random forest analysis in this case better confirms the results from the single tree. In this analysis, again the baseline measure is the most important in the permutation importance statistic, with a value of .027. The next most important measure is the count of trauma symptoms with an importance score of .012. Education level also appears as potentially important, with a score of .01. The rest of the measures are quite close to zero or negative.

Because the outcome in this analysis is categorical, we can calculate the extent to which the random forest ensemble correctly predicts follow-up housing status. As described in the Methods section, two types of predication are available, the prediction for the overall sample, and the prediction for the "out-of-bag" cases only. For the overall sample the model predicted 78% of the cases correctly. When restricted to the out-of-bag sample, which would be a better indicator of how the model would do predicting follow-up housing status for a new group of women, or

for women in a slightly different context, the prediction accuracy drops to 57%.

3.2.5 Provider Feedback

The results of the analyses were presented to four staff from Project Genesis and their views obtained in a 2-hour focus group format. Staff attending the focus group included two front-line Care Coordinators, the Project Director, and the Agency's Director of Outpatient Services, the administrative unit in which the program was housed.

The staff were asked five broad questions. First, they were asked, before seeing the analysis results, their own views on the differential impacts of project Genesis, of whether they felt their program had been more or less successful for different types of clients. The group generated several responses to this question. Table 3.8 shows the factors they identified as possibly driving outcomes.

The providers in general reacted positively to both sets of results, and were able to understand and discuss them in depth. After they had reviewed the results, the providers answered questions concerning the understandability, scope, validity, and usefulness of the different methods in a general discussion. At the conclusion of the group, each provider separately completed a short questionnaire (see Appendix C) in which she used a semantic differential format to summarize her views on the four questions and an additional item asking which method she would prefer for obtaining regular program updates. The results of these ratings are shown in Figure 3.6.

In terms of the understandability of the methods, the providers felt that both methods were understandable with explanation, and that neither would be without explanation. Within that overall pattern, there were differing opinions on the understandability of the trees. In the discussion, one provider felt that "the re-

Potential Factor	Description
Street dwelling vs. shelter dwelling	Women living on the streets had more severe conditions and situations, and were probably less likely to show large, easily measurable positive outcomes such as getting housed or becoming sober. These women may have been helped in more subtle ways such as learning new coping skills. On the other hand, starting in more severe conditions, they also had more room to improve.
Social class of family of origin	Some women were from middle class backgrounds and had broader experiences of ways to live than women who had grown up with inter-generational poverty. These women had more understanding of and access to middle class culture and norms which may have helped them exit homelessness more easily.
Referral source	Project Genesis recruited from multiple different types of agencies and this may have impacted how women perceived the program, and therefore what they got out of it. Women who were recruited from the large substance abuse treatment center or from homeless shelters saw the program as a defined entity. Women recruited via street outreach did not necessarily receive a broad overview of the program. Therefore, women recruited from the large institutions may have known more about the program, expected more, and understood what they could obtain more.
Partner status	Women who had male partners were less likely to succeed. The relationships women were in were frequently unhealthy and often violent. Men were often responsible for women losing housing, engaging in substance use and other risky behavior, and for emotional crises. It was easier to stabilize women who did not have partners. Women with partners could be stabilized, but crises were much more likely to occur and undo progress that was made.
Co-occurring disorders	Some women did not have significant mental health issues, and were dealing primarily only with substance abuse. These women, who had more "mental clarity", may have been more able to advocate for themselves and exit homelessness.

Table 3.8: Potential Factors Driving Outcomes According to Genesis Staff

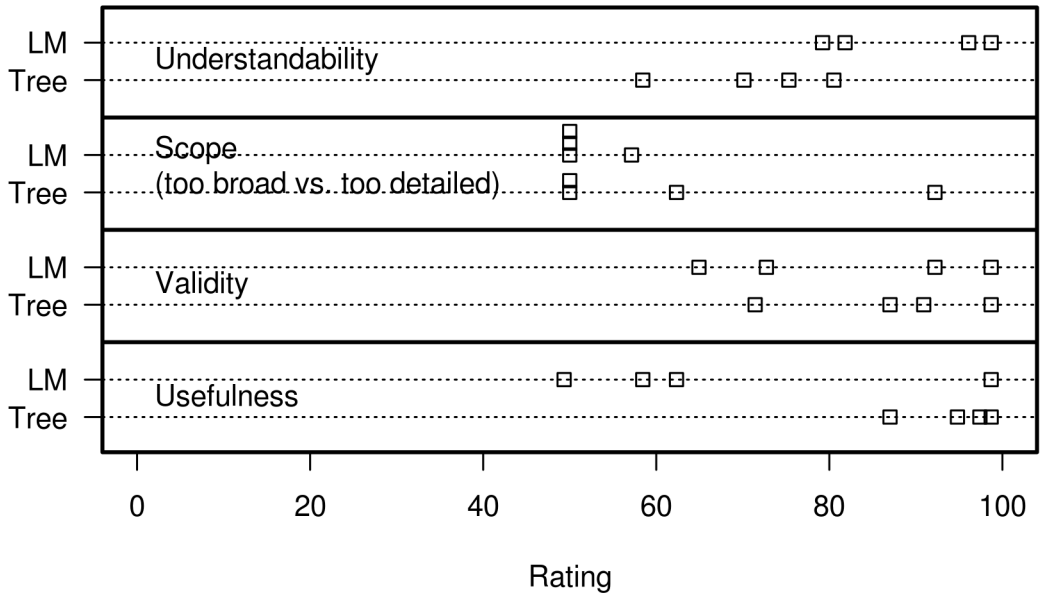


Figure 3.6: Ratings of Methods By Genesis Staff

gression is less accessible" but another that "you'd need to do more to get the trees across [to providers]". The visual presentation of the trees was seen as an advantage for that method: "the tree is easier to go back and reference later", and "the tree is easier to grasp immediately; the other is just a bunch of words on a page". There was some discussion of how the actual tree diagrams could be made less intimidating, especially since the ones presented are as shown here, with p-values and the "node numbers" that derive from the fitting software. In the numerical ratings, the providers rated both methods on the positive side of the scale, but clearly rated the trees as less understandable than the linear models.

For the question of "scope", whether the methods produced findings that were too broad or too detailed, most providers put both methods near the middle of the scale. This scale, unlike the others, was anchored by negative endpoints at both ends ("way too broad" on the left and "way too narrow" on the right), so ratings

in the middle of the scale indicate satisfaction with the output. This question, of the four, was the most difficult for the providers to answer. One pointed out that the scope question really depends on what the tree method finds, how detailed it gets in producing significant "splits".

In the discussion, both sets of results seemed believable to the providers. Three specific points were made: First, it was noted that the regression method found more significant results than the tree method, and this was seen as a failing of the tree method. Second, there were some findings that were initially surprising, such as the finding concerning one Care Coordinator who had substantially fewer clients literally homeless at follow-up, but these were seen as retroactively explainable. Third, the providers were surprised by what was not found. Although they felt the results obtained were believable for both methods, given the variables that were entered into the analyses, they were surprised that more significant differences were not found, for example that service contacts and trauma experience were not stronger predictors.

The providers saw important uses for both methods. The group came to a strong consensus that the methods were in fact complementary. They saw the linear models as being more useful at a broader level, for program planners or program advisory boards trying to get a sense of an overall population or trying to target programs, while the tree methods were more potentially useful at the micro level for program managers and front-line staff in closely monitoring a program over time. As the former Genesis director said about the trees "this is helpful... if we got the tree in real-time we would have said 'we need to do a substance abuse group'." The trees, in crossing predictor variables, gave more targeted findings that the providers appreciated: "this feels more targeted, it gives me something to go after... Looking at the trees gives more of an idea of where to target it. I

would have missed that just looking at the regression." Because the item on the questionnaire specifically asks about usefulness for ongoing monitoring "how useful do you believe the results would be in alerting you to the need for potential program changes...", the numerical ratings strongly favor the trees over the regressions. In the discussion however, the providers acknowledged an important role for the regressions at a broader level.

3.2.6 Summary of Findings

The results from the analyses of the two outcomes using both the standard linear (ANCOVA) models and the tree models are summarized in bullet form below.

In the linear model for psychological symptoms, symptoms at follow-up were related to age, homelessness duration, and recency of alcohol/drug use, controlling for baseline symptoms:

- Baseline symptom score is strongly related to follow-up score
- Older women had worse symptoms at follow-up than younger women
- Women with 2-5 years of homelessness had worse symptoms than women with less homelessness; but, women with more than 5 years homeless history *did not*
- Women with recent alcohol/drug use had worse symptoms at follow-up

The tree model for symptoms found a smaller set of predictors to be statistically significant predictors than the linear model found: only the baseline value of the outcome and the substance use measure. The forest analysis disagreed with the single tree analysis in finding trauamtic experiences as the most important predictor after baseline symptoms, rather than drug/alcohol use:

- The model made splits on baseline symptoms and recency of alcohol/drug use.
- The final grouping of women is:
  - Women with very low symptoms at baseline, who end up with low symptoms at follow-up
  - Women with mid-range symptoms at baseline who were not currently using alcohol/drugs, who end up with medium-low symptoms at follow-up
  - Women with mid-range symptoms at baseline who were currently using alcohol/drugs, who end up with medium-high symptoms at follow-up
  - Women with very high symptoms at baseline, who end up with very high symptoms at follow-up
- The random forest found baseline symptoms to be by far the most important predictor, followed by trauma experience, recency of drug/alcohol use, and homelessness duration

For the homelessness outcome, which it must be remembered represents literal homelessness, living on the streets or in shelter, the linear model found older women to be more likely to be homeless at follow-up as well as two more anomalous findings regarding service use measures:

- Women literally homeless at baseline were more likely to be literally homeless at follow-up
- Older women were more likely to be homeless at follow-up than younger women
- Women working with one Care Coordinator were much less likely to be homeless than those working with the other three; this is likely due to differences in where the Care Coordinators recruited from

- Women with more service contacts were *more* likely to be homeless at follow-up than women with fewer

The tree model for homelessness did not produce any overlap in predictors with the standard linear model. It found traumatic experiences to be the only important predictor after baseline status.

- The model made splits on baseline homelessness status and number of traumatic events experienced
- The final grouping of women is:
  - Women who were literally homeless at baseline, who were likely to still be that way at follow-up
  - Women who were not literally homeless at baseline with (relatively) fewer trauma experiences, who were unlikely to be homeless at follow-up
  - Women who were not literally homeless at baseline but who had extensive trauma histories, who were likely to be homeless at follow-up
- The random forest model found baseline housing status as the strongest predictor, followed by trauma experiences

CHAPTER 4

DISCUSSION

Having seen the results from the standard and alternative methods in both Study A and Study B, we can now evaluate the methods against the five original questions posed in the Introduction, and explore the inquiry's results in a broader sense.

4.1 Evaluation of the Alternative Techniques

The questions posed in the Introduction ask about the relative merits of the alternative techniques, as compared to the standard ones, on five different dimensions:

- Comparability of results
- Representational complexity of results
- Utility of results to service providers
- Validity, reliability, and defendability of results
- Data requirements

In this section I address each of these questions in turn, assessing the relative merit of the alternative techniques along these each dimension.

4.1.1 Comparability

The first question to be addressed concerns the comparability of the results obtained by the standard and alternative methods in each study. Besides being of interest in its own right, this question must be addressed first because if the standard and alternative techniques produced results that are not meaningfully different from one-another, much of the subsequent analysis is moot as two of the

four remaining dimensions on which the techniques are to be evaluated (their representational complexity and their utility) would drop away. In the Introduction I outlined four potential outcomes regarding the comparability of results that seem to cover the universe of logical possibilities. The results from the two approaches used in each study could be 1) similar, 2) different but conformable with each other, 3) different and unrelated to each other, or 4) different and contradictory to each other. The first question to be answered is "How did the results fall out in terms of these four possibilities?"

The goal of the analyses in a cross-program comparison like Study A is to determine whether one program is more effective than another, controlling for baseline differences. Looking back at the results form the HLM models and the MCS analysis, we see that on this critical question both techniques provided some evidence for there being more change in Site B than in Site A, but both also indicated that this change appears to be due to baseline differences in the populations at the two sites, and cannot be considered a difference in efficacy between the programs. This was found in the HLM model for substance use frequency and in the tests for the MCS: While the "raw" test between the sites on MCS was significant, as was the coefficient for site when the MCS was modeled with demographics, the site effect disappeared once the background factors were entered into the MCS predictive model. Therefore, we can conclude that in terms of the overall question "Was one site more effective than the other?" both methods have come to the same conclusion — "no".

To provide a more specific look at the methods' comparability in Study A, I summarize in Table 4.1 the modeling results from the study across both methods used. The first three columns show the results from the three individual HLM models for the three outcomes: Substance use frequency, emotional problems, and

health problems. The final column shows the results from the Maximum Change Score model. Recall that with the HLM analysis, a predictor might impact clients' starting place on the outcome of interest (their baseline value), or their rate of change over time, both of these quantities, or neither. In the Table, a "B" indicates a relationship between the predictor and baseline status of the outcome, while a "C" indicates a relationship between the predictor and change on the outcome over time. Plus and minus signs, where appropriate, indicate the direction of the effect. It is important to note that in the MCS model, with its simple pre/post framework, the relationships between the predictors and baseline status are not explicitly modeled, so the column for MCS cannot have any "B" results, only "C" results, indicting that the predictor is related to change. A cell with dashes indicates variables that were not entered into the model; the substance problems predictor was not entered into the model for substance use frequency as it would likely be collinear with the baseline value of the outcome, and the same is true for the mental health predictors with the emotional problems outcome. Also, the MCS model does not test for an overall rate of change.

Predictor	**Sub. Prob.**	**Emot. Prob.**	**Hlth. Prob.**	**MCS**
Age 21+			C–	C–
Female				
Nonwhite				
Overall Rt. of Change				—
Site				
MH - Internal	B+ C–	—		C–
MH - External	B+	—		C–
Sub. Problems	—	B+	B+ C–	
Trauma	B– C+	B+		
Crime			B–	

Table 4.1: Comparability of Study A Results

When we examine Table 4.1, and keep in mind that only the "C" results indicating relationships between the predictor and change are comparable across

the two sets of results, we see that the results, while not exactly similar, are not strongly contradictory with one-another. As just noted, the models agree on the key point of finding no site differences once other factors are controlled for. Many of the statistically significant findings in the HLM models concern relationships between background factors and starting levels of the outcome, for example clients with higher baseline levels of mental health issues having higher starting values of substance use and vice versa. These results are not really surprising, nor very interesting, and presumably could be obtained without complex modeling by simply examining bivariate relationships between the variables in the baseline data. Of the five predictors that show any relationship to change over time across both sets of analyses, the MCS and HLM models agree on those concerning age and internalizing mental health problems, while the MCS has a relationship to externalizing problems not found in the HLM models, and trauma and crime both have relationships not found in the MCS.

This lack of exact similarity in results, even though the MCS derives directly from the other three outcomes, raises an interesting point and may be considered a drawback of the MCS's composite nature. It is difficult to interpret the MCS's relationships to other factors, such as background conditions, because it is itself a composite. Because it is based on multiple measures, we don't really know what it is we are relating to other measures. The final MCS model finds that clients with higher background mental health problems had greater change on the MCS. Since the emotional problem score was the most frequent maximally-changing measure, this is probably just an artifact and is simply indicating that clients with more mental/emotional problems at baseline experienced the biggest changes in mental/emotional health. Unlike with the HLM models, where the "same domain" background factors were excluded from the models to avoid this

collinearity, that approach is not possible with the MCS because it is a composite across domains.

In the main question of the analysis, whether the sites differ, the results for the MCS and the HLM analyses fall into the "similar" category, but in the details there is more discrepancy, enough to qualify for the "different but conformable" category, or perhaps the "different and unrelated" category, but certainly not the "different and contradictory" category.

Turning now to the comparability of the results for Study B, we find less agreement. Table 4.2 summarizes the results from Study B. The first two columns show the results for the psychological symptom measure and the second two show the results for the literal homelessness outcome. Within each outcome, the first column shows the results from the linear model (ANCOVA model) and the second column shows the results for the Tree and forest analyses. In the table, a plus sign indicates a positive relationship between the predictor and the outcome, an "X" represents a relationship between a categorical predictor and the outcome (which does not have a positive or negative direction), and an "F" indicates a predictor identified as important in the random forest analysis.

	Psych. Symptoms		**Literal Homelessness**	
Predictor	**Linear**	**Tree**	**Linear**	**Tree**
Age	+		+	
Race				
Hispanic/Latino				
Education				
Care Coord.			X	
Svc. contacts			+	
Psych. admits				
Homeless dur.	X	F		
Drug/alc. rcncy.	+	+ F		
Trauma		F		+ F

Table 4.2: Comparability of Study B Results

Here we find that the methods disagree to a substantial extent. All four of the analyses (two variables times two approaches) find, as expected, the baseline value of the outcome to be strongly predictive of the follow-up value (not shown in the table). The baseline and follow-up assessments were only six months apart, a relatively short time-frame for making substantial changes in homeless women's lives, and the baseline values of the outcomes are naturally strongly related to the follow-up values.

On the psychological symptoms score there is reasonable agreement in the sense that the only non-trivial variable found to be important by the tree method, recency of drug/alcohol use, was also found to be important in the linear model. But, the linear model also found age and homelessness to be important, where the tree model did not. The random forest analysis backed up the single tree model in identifying drug/alcohol recency as important, so all three methods agree on that predictor. The forests agreed with the linear model in the importance of homeless duration, but also added traumatic experiences as important. On the models for literal homelessness there is no agreement: The linear model found age, Care Coordinator, and service contacts to be important while the tree model and the forest analysis found trauma experiences to be. It seems therefore that these results fall into the "different, and contradictory" category in the original conceptualization from the Introduction.

This discrepancy in findings might be due to the different techniques' relative strengths and weaknesses. The effects in the linear models represent the association of the predictors to the outcome, with the other factors held constant. So, the significant effect for age in the mental illness symptoms model indicates that older women have worse symptoms than younger women, if we remove any impact of the other variables on symptoms. By allowing all the two-way interactions to enter

the models we checked for relaxations of this stricture, but none of the interactions were significant. The tree models on the other hand take a different approach and must look for important variables within the groups defined by the previous splits that have already been made. Thus, it is very difficult for the tree models to find a non-interactive effect. If the first split made is on variable A, then to find a non-interactive additional effect of variable B, the tree algorithm would have to find that variable B is the most important predictor within each of the two groups formed by the split on A, and it would have to find that the specific value of B for splitting on was the same in both groups. This is clearly an unlikely set of events. So, the linear model method found main effects, but no interactions entered the models, while the tree models found only interactions, and no straight main effects. It is puzzling however as to why in the homelessness model the variables identified by the two had no overlap. This pattern could arise from some variables being important in interaction with others — that is only for subgroups — which would be picked up by tree models but not necessarily found by the linear models where power to detect interactions is typically lower than power to detect main effects.

4.1.2 Representational Complexity

The second dimension for evaluation is the representational complexity that the techniques provide. As outlined in the Introduction, the motivation for this study is to find analytic techniques that might better portray the richness and complexity inherent when complicated clients interact with complicated human service programs. Turning first to Study A, there are two overriding points that argue for, and against, the benefit of the maximum change scores in this regard. First, it appears that the maximum change score is generally functioning as hoped in providing increased flexibility. When the three different variables are combined,

meaningful numbers of clients end up having their largest change on each of the three measures. This was in no-way pre-ordained and is an important finding. It indicates that there is between-client variability that is potentially important and worth capturing. If a great majority of clients ended up changing the most on just one of the measures — psychological symptoms say – then the basic premise of the MCS's potential to increase representational complexity, that it allows different clients to be impacted differently by a program, would be vitiated. This did not occur in this case. The strength of the MCS is that it is to some extent a *person-centered* rather than *variable-centered* measure (Magnusson & Torestad, 1993): It conveys some information about the arrangement of characteristics *within* the person, as opposed to a standard measure which only indexes differences *between* people. It appears that in this case anyway there is some meaningful information in the MCS.

The second point, which argues against the MCS's representational complexity compared to that provided by the hierarchical models, concerns its flexibility for modeling. The MCS is not defined for multiple timepoints. It would not necessarily make sense to calculate it, for example, as change from baseline to 3 months and change from baseline to 6 months, since for any given client the measure they were changing on might be different between the two pairs of timepoints, and therefore not comparable. Harking back to the three types of longitudinal analysis described by Hedeker and Gibbons (2006) and Fitzmaurice et al. (2004), the MCS is a "summary measure" — it collapses change over time down to non-longitudinal measure. Because of this, the MCS is only really defined for pre/post situations or assessing change from a pre-test to the latest follow-up. It is not amenable to full longitudinal analysis as provided in hierarchical models. One of the chief benefits of HLM is that it allows the researcher to examine explicitly and in a single

model factors that predict starting status alongside factors that predict change over time. This dual capability increases the representational complexity of HLM since we can tell a richer story about clients that maps directly onto the real world (for example "clients at site B started higher on psychological symptoms, but showed sharper declines"). Because the MCS collapses the time dimension into itself, we are deprived this more complex and useful form of modeling.

To summarize, the MCS worked as expected — it allowed different clients to be impacted differently by the program — thus increasing the representational complexity of the analysis. But, because it is a summary measure across time, we lose the capability to portray starting place and change over time, a powerful and natural way to analyze longitudinal data and tell a rich story. Therefore, I conclude that the MCS does not offer higher representational complexity than the standard HLM approach.

In Study B, as with Study A, there are two main features in the results obtained that work in opposite directions. On the one hand, the propensity of the tree models to find interactions rather than main effects (described above) did, as expected, provide for more nuanced, detailed findings than the linear models. The tree model for psychological symptoms identified four groups that differed on the outcome, formed by crossing baseline symptoms and drug/alcohol use. The tree model for literal homelessness found three groups that differed on the outcome, formed by crossing baseline homelessness status and traumatic experiences. The framing of results into smaller and smaller sub-groups in these models naturally leads to a more richly textured set of findings. This approach provides easily interpretable findings in a way that linear model results, with their notion of "partialing", "controlling", "removing", or "holding constant" other effects, are unlikely to match. Furthermore as Berk (2004) points out in his interesting critique

of the use of linear models, there are many situations where it may be inappropriate to consider marginal effects of a predictor while holding constant other variables in the model. For example, in the linear model for psychological symptoms there was an effect for age, indicating that older women had higher levels of symptoms at follow-up, controlling for baseline levels and holding the other factors such as race, education, and duration of homelessness constant. But, it is quite plausible that in the real world these other factors do differ by age, so how important is the marginal effect of age? Berk reminds us that to interpret such a finding we have to at least hypothetically be able to imagine that the age of a representative woman in the sample changes, while nothing else about her does. The linear models, in attempting to provide us with the predictors' independent effects, may induce a degree of artificiality into the analysis. The tree models make no claims to disentangle the effects of predictors, and simply tell us that, say, the group of women with medium symptoms at baseline who were recent drug/alcohol users did not do well in terms of symptoms at follow-up. This type of result, while more qualified and perhaps less generalizable, provides a richer, more nuanced picture than the main effect results obtained in the linear models.

On the other hand, the representational complexity of the trees was hurt by the shallowness of their branching, that is by the relatively few non-trivial splits that the models identified. In each tree model, there was only one split found that was not simply a split on the baseline value of the outcome. While this may well reflect reality, it was disappointing from the standpoint of showing a rich, diverse, interesting set of findings. The standard models were able to identify more factors as statistically significant predictors. So on the one hand, the linear models found more variables to be important, which seems to tell a richer story, but on the other hand they did not find any interactions, so the story they tell

is in that sense less rich and nuanced. The low number of splits found by the tree algorithm is due to the stricter statistical standards used in this modern version, conditional inference trees, than in the classical CART approach (Hothorn et al., 2006; Strobl et al., 2009). While classical CART trees can be run fully until all the subjects are in homogeneous or singleton groups, the conditional inference trees use statistical tests with corrections for multiple tests at each step to determine whether a statistically significant improvement in fit is produced. This approach addresses the main criticism of trees, that they overfit, but in doing so may overcompensate. It may be possible to relax this criterion by setting the p-value for splits somewhat higher, and then seeking validation of results in an ongoing evaluation context by repeating the analysis as new data comes in, or by conducting more in-depth examination of trees produced through a random forest analysis (see the potential use of Hasse diagrams below). Though it is a matter of judgment, it seems to me that the benefits of the trees in finding interactions and framing the results as based on sub-groups outweigh their disadvantage of finding fewer predictors, and it appears that the trees do offer increased representational complexity over the standard models. This is particularly true when one considers the point of view of service providers, which is the next dimension discussed below.

4.1.3 Utility for Service Providers

An important component of Study B was assessing the value of the results obtained via the different techniques from the point of view of the service providers who were involved in the project. As described in the Results section for Study B, the service providers' view of both methods were largely positive and emphasized the complementarity of the methods over the superiority of one method or the other. The group found both methods to be understandable, with explanation,

but struggled somewhat more with the tree models than with the standard models. It was clear from the discussion that two of the four providers grasped the tree models completely, but it was unclear how much the other two understood. As an upside for the tree approach, the group found the graphical presentation very appealing. There were no strong feelings that one set of findings was more valid than another — the staff expressed surprise that more significant findings were not obtained, particularly for the trees, but did not feel that any positive findings were blatantly erroneous. The group, which included front-line case managers (the Care Coordinators) as well as more senior staff, believed both methods could help them in their work. They saw the linear models as providing a broader, more overview-level set of findings, while the trees provided a more targeted, micro-level set of findings. The former might be of more use for program planners while the latter would be of more use for ongoing program monitoring. The distinction between broad, main effects for the standard models and narrow, interaction-based effects for the tree models arose organically from the service providers' discussion and reflects, I believe, an important point and the main reason the trees can provide higher representational complexity than the standard methods, as I argue above. The service providers appreciated the detailed sub-group based interpretation provided by the trees. One comment, from the program director, illustrated this particularly. She said "this [the tree model] feels more targeted, it gives me something to go after . . . I would have missed that just looking at the regression."

4.1.4 Validity

This study does not provide a rigorous framework in which to evaluate the validity of the results obtained by the different methods. Since the study draws upon real data, in which the actual patterns are unknown and partially discoverable

only through the application of the various imperfect methods, we lack a "gold standard" against which to judge which method had produced the more accurate, representative, or valid results. This situation could be changed through the use of simulation methods and a study along those lines is suggested below as a direction for future work.

We can, however, provide some "softer" evidence for the validity of the results. One important piece of this evidence was described above: The service providers, while surprised that more factors were not found as predictive in the models, felt that all the results obtained had a high level of "face validity". They believed all were valid, interpretable, and sensible given their experience with the program.

For the Maximum Change Score method, the fact that the raw change scores going into the MCS are analyzable, plotable, and interpretable on a common scale greatly increases one's confidence in the technique. As Boothroyd et al. (2004) illustrated and I followed, it is possible to analytically track the construction of the MCS. One can examine the input change scores, compare them to each other, and it is not a great leap from them to the final MCS. In this regard, the construction of the MCS is intuitive and sensible, and its interpretation as "the measure the client changed the most on", is straightforward.

On the other side, the complexity of the HLM analysis to some extent detracts from its apparent validity. There are so many parameters in the HLM models, so many ways the models could be parameterized, and so many potential ways parameters could be tested *post-hoc*, that it is difficult to come to conclusions. For example, in the emotional problems model, there was the interesting situation where the base rate of change coefficient, which represented the change for young white males in Site A, was not significant. This indicates that this group did not change over time in emotional problems. The coefficients for site and demographic

variables predicting rate of change were also not significant, indicating that being female, being nonwhite, being older, or being from Site B did not significantly add to nor detract from that non-significant rate of change. However, if one tests the slope for young white males in Site B directly using a *post-hoc* test, it is significant, indicating that young white men in Site B did have a significant decrease in emotional problems, unlike those in Site A. While the amount of their change was not significantly different from the amount of change experienced by young white men in Site A, it was enough, when added to the base amount, to be significant overall. It may be possible to parameterize the model differently to better account for such situations, perhaps by "deviation coding" where there is an overall rate of change, and then deviations of groups (by gender, race, site, and background factors) from this overall rate. The problem is, that there are potentially so many ways to parameterize the models that it can seem arbitrary to choose one particular constellation of choices over another. Crawley (2002, p. 431) makes this point decidedly in discussing standard linear models: "The thing to remember about multiple regression is that, in principle, there is no end to it. The number of combinations of interaction terms and non-linear terms is endless". This situation applies even more to HLM, where many further decisions need to be made beyond those for a standard model. HLM is both tremendously more complex than the MCS technique, and also tremendously more flexible. It can no doubt yield very solid, reliable results, but at a very high cost in terms of expertise, time, and effort.

In Study B there are both positive and negative aspects to the validity of the tree models. On the one hand, the results seem intuitively valid and reasonable. In both tree models, extreme values on the baseline measure resulted in extreme values on the follow-up measure. In the mid-range, where there was perhaps more "causal space" for other factors to come into play, each tree model found

one very plausible predictor. On the downside for the trees, however, was the discrepancy between the tree and forest findings for the mental illness outcome. The forest found trauma experiences to be the most important variable (after baseline value of the outcome), while the single tree only included alcohol/drug use besides the baseline outcome value. As the forest analysis is based on averaging across 1,000 trees done on random subsets of the cases and the variables, the importance measure it generates must be given some credence. This may be a case where the substance abuse measure has a stronger relationship with symptoms when considered near the top of the tree, but that trauma experience, interacting with more variables further down the tree, is overall more important but cannot enter the single tree model because it is "beaten out" by substance use at higher levels. So traumatic experience, if it interacts more with other variables, might have a larger overall impact than substance use but substance use enters first in the single tree because its effect is more concentrated. For the homelessness model, the forest analysis supported the single tree model, which is reassuring. The overall predictive accuracy of 78% for the forest analysis of homelessness seems strong, but the "out-of-bag" prediction accuracy of 57%, which is a more reliable measure because it is not using the data used to actually make the trees, seems rather weak.

4.1.5 Data Requirements

Neither alternative technique poses substantial problems in terms of the requirements it makes of the data, or of the evaluation context, with two exceptions for the MCS. One of these restrictions, that the MCS is not defined across multiple timepoints, is described in the discussion of representational complexity above. Another restriction of the MCS is that it is currently only defined for continuous outcomes. In the construction of the MCS, each continuous change score is normal-

ized by the variability of its corresponding baseline outcome and then combined with the parallel scores from the other outcomes of interest. This procedure would not allow for a set of outcomes that included both continuous and categorical outcomes, or for a set of purely categorical outcomes. By developing a somewhat more complex weighting scheme, it seems possible that this hurdle could be overcome and information from multiple types of variables might be combined. This would involve some probably arbitrary weighting of change on the categorical variables compared to change on continuous ones, but could certainly be developed if scaled against some conception of what represents a clinically significant change in each instance. Within these constraints of only working on two time points and requiring continuous measures, the MCS does not pose any further restrictions, as it derives directly from other outcome measures.

The tree models in Study B are extremely flexible and make no assumptions about the distribution of the outcome measures or of the predictors. Furthermore, the tree models can accommodate any number of predictors and are not bound by restrictions on the number of variables or parameters based on the number of cases available for analysis.

4.2 Limitations of this Analysis

The chief limitations of this study stem from its narrowness, as defined in a host of ways. It is a small trial of these techniques on two specific evaluation datasets, and cannot directly address the validity or utility of the techniques more generally. Since the results obtained with each of the different methods could vary with different outcome measures, different samples, and different interventions, the comparison of the methods to each other could come to substantially different conclusions in replications on other datasets. This is especially true for the tree

models, where the technique is especially sensitive to particularities in the data, since it decides variable importance sequentially, one split at a time.

There are at least five dimensions on which this study's specific nature might limit the generalizability of the results. First and perhaps most importantly are the particular interventions studied. Different types of interventions might produce results more suited to the standard vs. the alternative techniques. Simpler interventions with broad "causal granularity" (see Introduction) might best be modeled with standard models, whereas the alternative techniques might be better suited for small-granularity interventions. As noted above, the standard and tree models for example are differentially suited for finding main effects vs. interactions, so a more coarsely-grained intervention might actually be better suited by a standard model and a more finely-grained one by a tree model. Second, to make the analysis more tractable, I limited the Study A analysis to two of the fourteen AAFT sites. It would be interesting to conduct a MCS analysis across the entire cohort of sites. Are there AAFT sites in which most of the clients had their greatest change on one particular outcome? Does the finding of a similar makeup in the MCS across the two studied sites hold up when all fourteen are considered? Do some sites stand out among the entire cohort as providing more overall change on the MCS than others? These are fascinating questions and highlight that the results obtained for the MCS may well have been strongly determined by the particular characteristics of the two sites chosen.

A third and a fourth dimension are also related to the particular interventions studied: The particular outcomes used for analysis and the timescale over which they were assessed. In both the AAFT and Genesis evaluations, as is common in human services evaluations, multiple outcome domains were assessed because the program was expected to potentially have a broad or varying impact and/or there

was no strong program theory to limit the assessment to particular domains or measures. While I made selections among the variables available based on their overall importance to the interventions and their characteristics, it is likely that different patterns of findings would have been obtained with different measures. There are potentially complex and important causal interactions among outcomes with some serving as platforms for others. In the homelessness literature for example, housing stability is frequently cited as powerfully affecting other domains such as mental health, and substance use is frequently noted in a reverse pattern of adversely affecting residential stability (United States Interagency Council on Homelessness, 2010). An MCS analysis, by subsuming multiple outcomes into one measure might fruitfully reveal such interactions if the modeling of it with predictors could be done in a sensitive enough fashion. On the other hand, the composite nature of it might wash out subtleties that could be determined through careful longitudinal modeling of outcomes allowing other measures to enter the model as time-varying covariates. The short timescale for the results, particularly for Study B, might also have strongly shaped the results obtained. If women in Study B had more time for program-induced changes to take hold, we might have seen a very different pattern of results, perhaps resulting in richer "splits" in the tree models. It is possible that the results obtained were in fact only the embryonic stage of a pattern that developed over time for participants and would have been more fully revealed if we had 12, 18, and 24-month follow-up data on women from the program.

A final dimension concerns the focus and robustness of the particular interventions themselves. Both of these interventions, but especially the AAFT project, were relatively recent. The AAFT funding for the entire cohort was provided for only three years. While the A-CRA/ACC model that was employed in the project

has some evidence-based support (Dennis et al., 2004), the grantees implementing the model in this intervention had limited time to hire staff, have the staff get trained and certified, and build the required infrastructure to deliver the services effectively. Qualitative data indicated that some of the AAFT sites struggled with this process more than others (Tobin, Lang, & Huntington, 2012), with one of the two sites studied falling somewhat into this category. However, it is possible that the relative lack of site differences and change over time found in Study A was due to neither site implementing the intervention very robustly. The Genesis intervention was funded for five years so it suffered from this potential drawback less, but nevertheless its very open, global nature precluded it ever developing a manualized, robust set of standards for its practice. If the study had been conducted in the context of a service intervention with a longer history and more established practices, it is possible that more robust, and/or more differentiated results might have been obtained.

This point raises the interesting question of whether more rigid, defined, targeted interventions might have simpler patterns of causal granularity than more individually tailored ones, and therefore be better served by standard methods as opposed to alternative ones. While both the AAFT project and the Genesis project paid some lip service to tailoring their services to clients' needs, the Genesis project did so to an extreme extent — indeed there was little defined in the program's approach beyond "meet women where they are". Genesis in this regard contrasted with AAFT which had a robust set of manuals, trainings, and fidelity checks. With a program like Genesis, there is in fact no singular "program" to be studied since services are so highly configured to individual clients' needs. This feature, while likely beneficial for clients, makes evaluation research more difficult, and necessitates collection of detailed data on services received to even describe

what "the program" was for different clients. Brown et al. (2009) note that standard randomized trials are ill-suited for studying complex programs that adapt in response to client needs or preferences and that "adaptive" designs in which the study and possibly the intervention adapt in reaction to client status and/or group-level findings are becoming more common in medical and public health research (Brown et al., 2009). It's possible that the MCS could be used in adaptive designs, providing another method for accommodating client-level differences.

Another limitation of this study, related to its being a trial of the methods on two particular datasets from two particular programs was discussed above under Validity: we have no external basis for knowing the true state of affairs in either Study A or Study B. While we are comparing the results from two different methods to each other, neither may be a very good reflection of the reality present in the data.

One powerful solution to both of these limitations is to expand from a simple trial of the techniques on real-world data, as was done in this study, to a simulation study in which artificial datasets, with known characteristics, are created and the techniques are assessed on those data (see proposed approach discussed below). In fact, Strobl et al. (2009) explicitly calls for such simulation studies for the tree based methods, noting that repeated trials on idiosyncratic datasets may produce diminishing returns in terms of moving the field ahead.

4.3 Overall Evaluation and Directions for Further Research

Given what we have found regarding these methods' comparability, representational complexity, utility, validity, and data requirements, what are the implications of these results for the larger attempt to better capture the rich interplay of clients and programs in human services evaluation? As Berk (2004) has noted,

there is certainly room for improvement in applied social scientists' use of statistical modeling, particularly when questions of causality such as program efficacy are in play. What do these results say about improving analytic techniques to model the complexity inherent in human services evaluations?

A summary of the evaluation of the alternative methods on the five dimensions from the Introduction is provided in Table 4.3. Given the results, it seems that this demonstration has provided evidence that both of the alternative techniques can be potentially useful in particular human service evaluation contexts, but that the tree models are likely more so than the maximum change scores, without further development. There are four shortcomings of the MCS, that do not directly affect its suitability for the job for which it was designed, but simply limit the contexts in which it can be applied. First, it can only be used in group-comparison situations where the fact of it being biased (overstating the true amount of change) is offset because the same bias is operating in both groups. Second, the difficulty in interpreting its relationships to other measures, because it is a composite, is also limiting. It can summarize information across multiple variables, but it is difficult to interpret its relationships to other factors. Third, in its present form, it is a simple pre/post change score measure, and cannot be modeled in a richer longitudinal framework using HLM. Finally, in its present form, it is defined only for continuous measures and could not be used to summarize information across outcomes of different measurement levels.

The tree models provided limited, but believable and seemingly valid, results in Study B, which the service providers saw as useful. These results differed significantly from those obtained through the standard linear model approach, and this is likely due to the trees' propensity to detect interactions over main effects. The ability of the tree techniques to work on continuous and categorical outcomes, as

Dimension	Study A: MCS	Study B: Trees
Comparability	Similar on overall main question; different but conformable in the details.	Different and contradictory.
Rep. complexity	Worked as expected, but cannot be modeled with sophisticated longitudinal techniques.	Found interactions and subgroups and therefore a richer picture, but fewer predictors overall.
Provider view	(Not assessed)	Positive view of both methods; appreciated detailed findings and visual aspect of trees; felt trees useful for program monitoring.
Validity	MCS construction transparent and believable vs. HLM models that are powerful but exceedingly complex, difficult to implement.	Pattern of results for both variables sensible and believable, but trees and forests did not perfectly agree.
Data reqs.	Not defined for multiple timepoints and categorical measures.	Extremely flexible, no limitations to use.

Table 4.3: Summary of Evaluation Dimensions

might be found on many clinical assessment tools, is important.

These results suggest that future methods development work could fruitfully focus on expanding the flexibility of the MCS. In particular, expanding the technique to allow multiple timepoints and different measurement levels would be ideal. The first task might be accomplished by defining each client's maximally changing measure over the full range of time assessed, and then examining change between baseline and intermediate follow-up points on that measure. Besides these specific expansions, there is room to define the MCS in other ways to capture different aspects of change. Once standardized individual change scores are calculated, one could in principle combine them in several ways such as taking the sum to get a "most change across domains" measure, or the count of domains meeting a certain threshold for change, or perhaps scaling them not statistically but in terms of

clinically meaningful benchmarks.

The tree methods could fruitfully be extended by finding a way to better summarize and visualize the results of the random forest analyses. The service providers reacted very positively to the visual results of the tree models, but the more reliable forest analyses can provide no visual output, only a variable importance measure that is scale-less and interpretable only within each study. Tree model practitioners recommend visually inspecting multiple trees from a forest to get a sense of the findings (Breiman, 2001; Strobl et al., 2009; Hothorn et al., 2006), but it would be helpful if there was a more systematic way to address this issue. As mentioned above, Hasse diagrams (Wikipedia, 2012; Brggemann, Patil, Brggemann, & Patil, 2011) might conceivably provide a way to visualize results from a forest analysis. These diagrams provide a way to visualize partially ordered sets of items, and the relationships among them. For example Bloch, Carter, and Cox (2012) constructed a Hasse diagram portraying the overall ranking of countries in the 2012 Olympics, using a set of rules incorporating counts of gold, silver, and bronze medals. Some countries were unequivocally better than others no matter how you counted or weighted the medals won, while for other pairs of countries there was no unambiguous relationship between them because their relative ranking depended on how you counted or weighted the different medals. These relationships were reflected in the Hasse diagram. The variables in a forest analysis, if considered across all the trees in the forest, might be seen as a similar partially ordered set of items where some are clearly below others, and others are indeterminate. A Hasse diagram approach might be helpful in visualizing this. Another potentially fruitful approach would be to use the predicted group membership from the forest analysis and retroactively profile the groups obtained on the measures that went into the forest. This method would be similar to the way

in which one interprets a cluster analysis by profiling the clusters obtained on the variables used to form the clusters (Kaufman & Rousseeuw, 2005).

A further potentially fruitful area for methodological developments is in the graphical display of statistical modeling results for non-researchers. Given the providers' positive reactions to the graphical tree displays, it would be helpful to develop methods for graphically displaying other modeling results, perhaps with real-time interactive capabilities on a computer screen. One can imagine monthly feedback sessions between evaluators and program staff in which the staff are able to interact with the program data visually. Such sessions might engage providers in the evaluation, provide important formative information, and lead to better data quality throughout the evaluation as providers come to see and learn from data. Dynamic graphical displays of statistical information are rapidly developing (Cortes, Pregibon, & Volinsky, 2012; Swayne, Cook, & Buja, 1998), and harnessing some of this power for human services evaluation could be an interesting and important area for future work.

A final potentially interesting area for further work would be in pursuing a quantitative measure of causal granularity. One could develop a multivariate client similarity measure based on a set of background measures at baseline. This would result in a half-matrix of clients by clients, where the numbers represented how similar each pair of clients was across the multiple measures. One could then calculate a corresponding matrix of similarities of the clients' outcomes at a follow-up point. By calculating the degree of similarity in outcomes, for varying degrees of similarity at baseline, one might be able to plot a curve whose slope would approximate a measure of casual granularity, the extent to which clients have to be exactly similar to have similar outcomes. This would probably fail in practice, but it would be interesting to try.

Besides these methodological developments, the findings from this study suggest four specific follow-up studies or projects that could be conducted immediately without further methods development work.

First, simply replicating the study on other human service evaluation datasets would have some value. As noted above, researchers in mental health and substance abuse services have noted that we lack an ability to predict which interventions will work well for which clients (Uher, 2011; Drake et al., 2008; Roffman, 2011) and perhaps the tree-based methods of Study B would better address this issue, at least if studies within a fairly narrow domain were chosen for replication.

Second, as mentioned above, it would be interesting to run the MCS analysis on the full cohort of AAFT sites and characterize the properties of the measure across all fourteen of the sites. This would give stronger information on the utility of the MCS as a person-centered measure in a broader set of contexts.

Third, and most importantly, it would be helpful to conduct a simulation study to evaluate the alternative and standard techniques' ability to uncover known patterns in artificial data. To conduct such a study following up on the tree methods in Study B, one could generate simulated outcome datasets that vary along two dimensions. The first dimension would be the strength of the underlying pattern in the results. This could be thought of as the signal-to-noise ratio of the pattern against the background of random variation in the data. It is possible to imagine a very strong differential outcome pattern, such as all women improve more than all men on a particular outcome, or women as a group have change scores that are two standard deviations larger than men's. On the other hand, one can imagine a very weak pattern in which there is a slight tendency for women to improve more than men. The second dimension would be the complexity of the underlying pattern. A pattern that simply involves one group defined by a

dichotomous characteristic improving more than the other (e.g. women improving more than men) is very simple. It is possible to imagine much more complex outcome patterns involving interactions of multiple characteristics, for example one in which clients in general showed improvement on an outcome measure of quality of life but the degree of improvement for women who came into the program with extensive trauma histories and lacking social support was markedly lower than that for other groups. This pattern involves a three-way interaction among gender, social support, and traumatic experience, and is more complex than the ones uncovered in Study B. By building multiple datasets that systematically vary along these two dimensions of pattern strength and complexity, and analyzing the multiple datasets with the same techniques, one would be in a strong position to assess the relative strengths and limitations of the methods for identifying the patterns built into the mock datasets and in discriminating these patterns from other, spurious patterns that the methods might identify.

Finally, it would be exciting to develop the tree methods as an ongoing system for program monitoring. Driven in part by federal guidelines for electronic health records and funder demands to document effectiveness, human service providers are more and more collecting ongoing client-level service use and outcome data through administrative systems. The tree methods are, for the end-user, relatively easy to use if the statistical algorithm could be provided in a package that interacted with client-level data already being collected and stored electronically in a management information system. In such a setup, the end-user, whether an evaluator or a program staff member, would only have to select the measures of interest for a tree model to be produced with sensible default parameters. Regular program team meetings where the same tree models are produced in real-time using up-to-date client data could provide a powerful mechanism for formative evaluation work.

Evaluators and service providers could interpret and discuss the tree results and hopefully develop an ongoing "virtuous circle" in which information is collected, summarized and fed back to providers through this system, and used by providers to improve the program, target their services, and better serve their clients.

APPENDIX A

RESULTS PRESENTED TO FOCUS GROUP

Focus Group to Evaluate Two Different Analytic Techniques

Genesis Results on Two Housing and Mental Health Measures:

Outcome	Baseline	Follow-up	p-value
Shelter/street was predominant living location in past 30 days	51.4%	38.7%	.006
BSI Global Severity Index – Measures how distressed or bothered the client is by 59 different mental illness symptoms • Ranges from 33 to 80 • Scores of 64 or more are considered “severe distress”	Mean = 65.0	Mean = 61.3	.001

Question we want to answer:

Are we helping all the clients equally, or are there subgroups of clients that are being helped more or less than others?

Two methods for answering this question to be evaluated:

- Method A: Regression Models
  - See whether characteristics of women at baseline predict their outcome at follow-up.
  - Each characteristic is examined for its impact on the outcome separate from the other characteristics, they are “controlled for”.
  - Include women's baseline value of the outcome to take it into account (“control for it”).
- Method B: Tree Models
  - Start with all women in one group. Find the baseline characteristic that most cleanly splits women into low and high groups on the outcome measure.
  - Within each of the two groups formed above, repeat the process to build an upside-down “tree” of splits. At each “branching” of the tree, find the characteristic that best differentiates women on the outcome.
  - At the bottom of the tree, the resulting groups of women should differ from low to hi on the outcome.

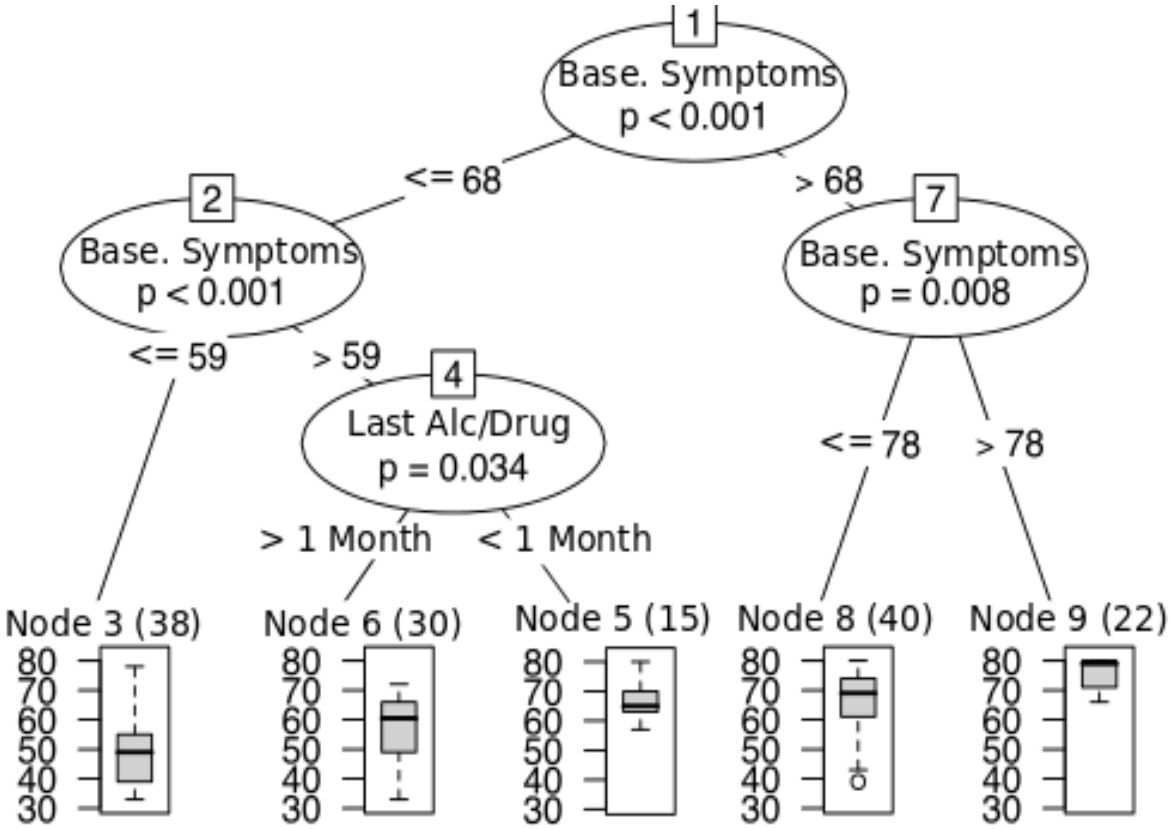
Approach to evaluate the two methods:

- I analyzed each of the two outcomes using both methods.
- I used the same baseline characteristics for each method:
    - Demographics:
        - Age
        - Race and Hispanic origin
        - Education
    - Services:
        - Care Coordinator
        - Number of service contacts
    - Severity of background conditions:
        - Ever been admitted for inpatient mental health (not used for predicting mental health symptoms)
        - Lifetime duration of homelessness (not used for predicting housing status)
        - Recency of last drug use
        - Number of different types of traumatic events experienced.

Results For Mental Health Symptoms
Method A: Regression

- Baseline symptom score
  - Of course, women with higher symptoms at baseline had higher symptoms at follow-up.
  - On average, having a score 10 points higher at baseline meant having a score around 7 points higher at follow-up.
- Age
  - Older women had worse mental illness symptoms at follow-up than younger women.
  - Ten years of age resulted in women having on average a 2 point higher score on the symptom measure.
- Duration homeless
  - Women who had lifetime homelessness of 2-5 years had higher symptom scores at follow-up than women homeless less than two years. On average 5 points higher.
  - Women who were homeless **longer** than 5 years, **did not** have worse mental health scores at follow-up.
- Recency of alc/drug use
  - Women whose last use was less than a month ago had higher symptom scores at follow-up, compared to women whose use was longer ago. On average 4 points higher.
- None of the other characteristics mattered:
  - Race, education, Care Coordinator, service contacts, traumatic events experienced.

Results for Mental Health Symptoms
Method B: Trees

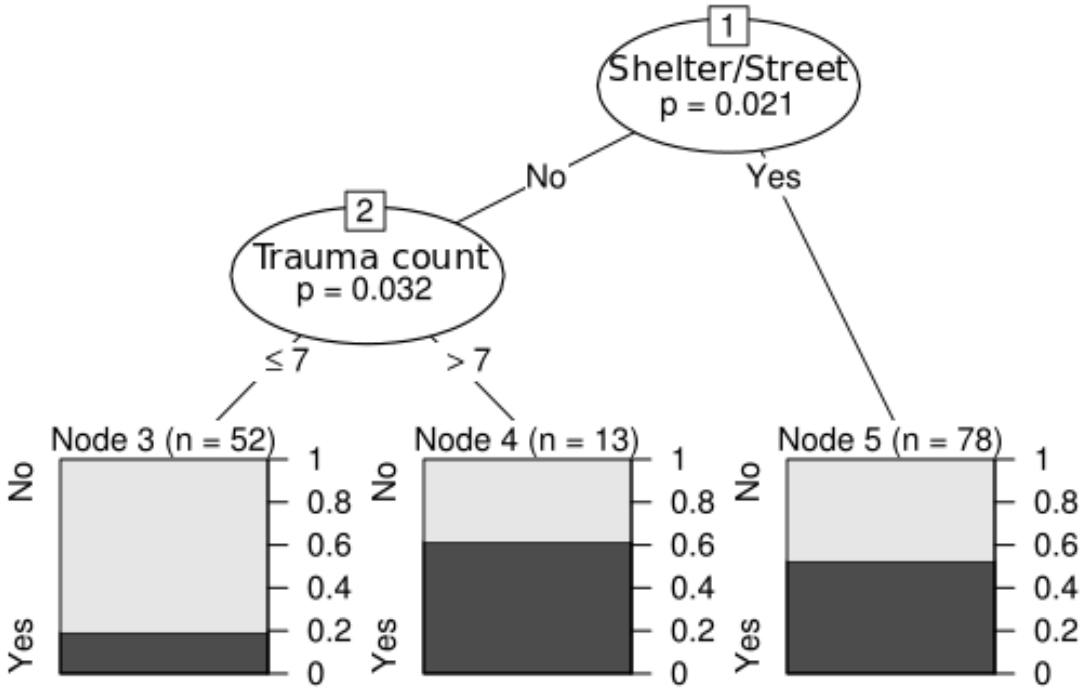


- As with Method A, the symptom score is strongly related to the follow-up score (as expected).
- The method first splits women into groups based on the Baseline symptoms. For women in the mid-range of baseline symptoms, substance use becomes important.
- The final groups are:
    - Women who had very low symptoms at baseline, who end up with low symptoms at follow-up.
    - Women who had mid-range symptoms at baseline who were not currently using drugs/alcohol, who end up with medium-low symptoms.
    - Women who had mid-range symptoms at baseline who were currently using drugs/alcohol, who end up with medium-high symptoms.
    - Women who had very high symptoms at baseline, who end up with high symptoms at follow-up.

Results for Literal Homelessness
Method A: Regression

- Literally homeless at baseline:
  - As expected, women on the shelter/streets at baseline were more likely to be literally homeless at follow-up, about three times more likely.
- Age:
  - Older women were more likely to be homeless at follow-up.
  - Wome ten years older were about 50% more likely to be homeless at follow-up.
- Care Coordinator:
  - Clients working with one care coordinator (Francesca) were much less likely to be homeless than those of others.
  - This is likely due to the different recruitment sources used by different Coordinators.
- Number of service contacts:
  - Having more service contacts was associated with **higher** probability of being homeless at follow-up.
  - This might be because women with access to housing were less likely to stay engaged?

Results for Literal Homelessness
Method B: Trees



- As with Method A, whether a woman was literally homeless at baseline strongly predicted whether she was at followup. This is the first split.
- Among women not literally homeless at baseline, trauma was important. Women with extreme trauma histories were much more likely to be homeless at follow-up than women without.
- The final three groups are:
    - Women who were not literally homeless at baseline with relatively low trauma experiences. Few of these women were homeless at follow-up.
    - Women who also were not literally homeless at baseline but who had extensive trauma histories. They were likely to be homeless at follow-up.
    - Women who were homeless at baseline were likely to be that way at follow-up.

APPENDIX B

FOCUS GROUP PROTOCOL

Analytic Alternatives Study: Focus Group Protocol

Introduction: Thank you for participating in this focus group to review results from the Project Genesis evaluation. As the primary service providers involved in Genesis, you have valuable insight into how the program worked, the people the program served, and the program's achievements.

Projects such as Genesis often serve clients with complicated histories and multiple issues, and naturally attempt to be as helpful as possible to clients, often providing a wide range of services and tailoring services and approaches to each client's needs, preferences, and situation. From an evaluation point of view, the interaction of complicated clients with complicated programs can present a challenge. Often the quantitative methods we have at our disposal are somewhat restrictive, and fail to capture the complexity of what goes on.

The purpose of today's meeting is to obtain your views on results from two different quantitative analyses of data from the project. Both analyses attempt to answer the question:

> Are we helping all the clients equally, or are there subgroups of clients that are being helped more or less than others?

One method has recently become the “best practice” for analyzing data from projects like Genesis. It is quite sophisticated from a statistical point of view and offers more flexibility than traditional techniques like regression analysis. The other technique was developed in another field and has not been widely used in human service evaluations, but may prove helpful. The goal of this study is to contrast and compare what can be learned using each technique. The hope is to make good tools available so that people working in programs can use data while a program is operating to quickly and easily examine whether they are helping all clients equally or whether there are sub-groups that are not being helped and may require more attention or a change in approach.

I have analyzed the data from Genesis using both techniques and am hoping to obtain your feedback today on the results. There are no “right” or “wrong” answers. With something as complicated as Project Genesis, each person involved might naturally have different opinions, based on their personal experience, about the specifics of what transpired. And since each of the front-line providers among you were working with different clients with your own personal styles and approaches, you will of course have had different experiences. I will be evaluating the two statistical techniques on some statistical criteria (e.g. are the conclusions statistically valid, do they make high demands of the data) but the goal of today is

to get your feedback as providers involved in the program who know better than anyone what the clients were like, what the program was like, and what you achieved.

You each have two short packets of results, one labeled A and one labeled B. I will review these with you and then I'm hoping we can have a brief focus-group discussion about your views of the results, and that you can each then complete a brief questionnaire as well.

Question 1: Before reviewing the results it would be helpful to get your own views on the question that these techniques are trying to answer. The techniques are limited by the data that goes into them and other factors, and may not produce valid results. Your views as the people who implemented the program, while they may reflect your own personal experience, are in some sense the "gold standard" to which the techniques should be compared. Do you think you were able to help all clients equally in Genesis or were subgroups of clients that you felt you were more or less able to help?

[REVIEW RESULTS PACKETS A AND B WITH GROUP]

Now that we have reviewed the results, I would like to ask you a few questions that attempt to get at four aspects of the results: their understandability, scope, validity, and potential usefulness.

Question 2: For any data analysis results to be useful, they have to be understandable. Statistical techniques vary in how comprehensible their results can be. Sometimes, statistical analysis produces results that are only really understandable by statisticians or people with extensive background or training.

Do you feel method A and method B each give results that you can understand?

Is method A or method B easier to understand?

Question 3: Statistical techniques can also vary in how much detail they provide. Sometimes results may seem too broad or general, and sometimes too detailed, or overwhelming.

How do you feel the results from method A and method B are in terms of the level of detail they provide?

Is method A or method B better in the amount of detail it provides?

Question 4: As the providers who implemented the project, your views on who was, and was not, helped by the project are likely to have validity. The statistical techniques may uncover valid patterns, or they may "uncover" patterns that arise by chance or are not reflective of reality for a number of reasons.

Do you feel that results from method A and method B are believable, have "face validity" to you, given your experience with the program? Do they "sound right" or "sound plausible" to you?

Does method A or method B provide results that seem more valid to you?

Question 5: The goal of this project is to identify tools that could be used to provide real-time feedback to people working in programs so that they could potentially alter their practice if that seemed warranted.

If you got results like those from method A and method B during [PROJECT NAME], do you think you could use those results to make mid-course changes to the program?

Do the results from method A or method B seem to be more useful, or more practical in terms of potentially making concrete changes to the program?

Question 6: We have attempted to cover four different aspects of these results, their understandability, their scope, their validity, and their usefulness.

Is there anything else positive or negative about either method A or method B that would like to add?

In general, do you have a preference for one method over the other?

APPENDIX C

FOCUS GROUP QUESTIONNAIRE

Analytic Alternatives Study: Summary Questionnaire

In this brief questionnaire you are asked to rate the two different sets of results on the same four dimensions we discussed during the focus group: understandability, scope, validity, and usefulness. For each question, please mark the horizontal line in the place that represents your view between each of the labeled end-points of the scale.

Question 1: How understandable were the results from Method A and Method B?

Method A	Not at all understand-able	|----------|----------|	Very under-standable
Method B	Not at all understand-able	|----------|----------|	Very under-standable

Question 2: How appropriate was the scope or level of detail of the results? (Please read the end-points of the scales carefully)

Method A	Way too broad	|----------|----------|	Way too detailed
Method B	Way too broad	|----------|----------|	Way too detailed

Question 3: How valid, correct or believable to do you feel the results from the methods are?

Method A	Very wrong, off-base, not believable	|----------|----------|	Very valid, on-target, believable

Method B	Very wrong, off-base, not believable	|----------|----------|	Very valid, on-target, believable

Question 4: How useful do you believe the results would be in alerting you to the need for potential program changes, if you obtained them during the program?

Method A	Not at all useful	|----------|----------|	Very useful
Method B	Not at all useful	|----------|----------|	Very useful

Question 5: If you could receive regular information on your current clients fed back to you using either method, would you have a preference?

Method A |----------|----------| Method B

Question 6: Do you have any other opinions not captured above about the two methods?

Thank you!

REFERENCES

Adamson, S. J., Sellman, J. D., & Frampton, C. M. (2009, January). Patient predictors of alcohol treatment outcome: A systematic review. *Journal of Substance Abuse Treatment*, *36*(1), 75–86.

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, *20*, 93–114.

Berk, R. A. (2004). *Regression analysis: a constructive critique.* Thousand Oaks, CA: Sage.

Bickel, R. (2007). *Multilevel analysis for applied research : it's just regression!* New York: Guilford Press.

Bloch, M., Carter, S., & Cox, A. (2012). The best and worst countries in the medal count. *The New York Times*.

Bonate, P. L. (2000). *Analysis of pretest-posttest designs.* Boca Raton: Chapman & Hall/CRC.

Boothroyd, R. A., Banks, S. M., Evans, M. E., Greenbaum, P. E., & Brown, E. (2004). Untangling the web: An approach to analyzing the impacts of individually tailored, multicomponent treatment interventions. *Mental Health Services Research*, *6*(3), 143–153.

Boulet, J., & Boss, M. (1991). Reliability and validity of the brief symptom inventory. *Journal of Consulting and Clinical Psychology*, *3*(3), 433–437.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1993). *Classification and regression trees.* New York: Chapman & Hall.

Bronfenbrenner, U. (1979). *Ecology of human development.* Harvard Up.

Brown, C. H., Ten Have, T. R., Jo, B., Dagne, G., Wyman, P. A., Muthn, B., & Gibbons, R. D. (2009). Adaptive designs for randomized trials in public

health. *Annual Review of Public Health*, *30*(1), 1–25.

Brggemann, R., Patil, G. P., Brggemann, R., & Patil, G. P. (2011). Partial order and hasse diagrams. In *Ranking and prioritization for multi-indicator systems* (pp. 13–23). New York, NY: Springer New York.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*, 297–312.

Campbell, D. T. (1986). Relabeling internal and external validity for applied social scientists. *New Directions for Evaluation*, *31*, 67–77.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Chen, H. T., Donaldson, S. I., & Mark, M. M. (2011, June). Validity frameworks for outcome evaluation. *New Directions for Evaluation*, *2011*(130), 5–16.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation : design & analysis issues for field settings*. Boston: Houghton Mifflin.

Cortes, C., Pregibon, D., & Volinsky, C. (2012). Computational methods for dynamic graphs. *Journal of Computational and Graphical Statistics*, *12*(4), 950–970.

Crawley, M. J. (2002). *Statistical computing : an introduction to data analysis using s-plus*. New York: John Wiley & Sons.

Dennis, M., Godley, S., Diamond, G., Tims, F., Babor, T., Donaldson, J., ... Funk, R. (2004). The cannabis youth treatment (CYT) study: Main findings from two randomized trials. *Journal of Substance Abuse Treatment*, *27*, 197–213.

Dennis, M., Titus, J., White, M., Unsicker, J., & Hodgkins, D. (2003). *Global appraisal of individual needs (GAIN): administration guide for the GAIN and related measures*. Normal, IL: Chestnut Health Systems.

Derogatis, L., & Melisaratos, N. (1983). The brief symptom inventory: An intro-

ductory report. *Psychological Medicine*, *13*, 595–605.

Drake, R. E., O'Neal, E. L., & Wallach, M. A. (2008, January). A systematic review of psychosocial research on psychosocial interventions for people with co-occurring severe mental and substance use disorders. *Journal of Substance Abuse Treatment*, *34*(1), 123–138.

Fitzmaurice, G., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, N.J.: Wiley-Interscience.

Forbes, A. B., & Carlin, J. B. (2005, May). Residual change analysis is not equivalent to analysis of covariance. *Journal of Clinical Epidemiology*, *58*(5), 540–541.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge; New York: Cambridge University Press.

Godley, S., Meyers, R., Smith, J., Karvinen, T., Titus, J., Godley, M., ... Kelberg, P. (2001). *The adolescent community reinforcement approach for adolescent cannabis users* (Vol. 4). Rockville, MD: Substance Abuse and Mental Health Services Administration.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, N.J.: John Wiley & Sons.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the A*, *81*, 945–970.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New: Joh.

Hothorn, T., Hornik, K., & Zeileis, A. (2006, September). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674.

Ives, M., Funk, R., Ihnes, P., Feeney, T., & Dennis, M. (2010). *GAIN: global*

appraisal of individual needs evaluation manual. Normal, IL: Chestnut Health Systems.

Kaplan, L. (2008). *The role of recovery support services in recovery-oriented systems of care: A white paper*. Washington, DC: Substance Abuse and Mental Health Services Administration.

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data: an introduction to cluster analysis*. Hoboken, N.J: Wiley.

Kessler, R. C., Chiu, W. T., Demler, O., Merikangas, K. R., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the national comorbidity survey replication. *Archives of General Psychiatry*, *62*, 617627.

Magnusson, D., & Torestad, B. (1993). A holistic view of personality: A model revisited. *Annual Review of Psychology*, *44*, 427–452.

Mark, M. M., Henry, G. T., & Julnes, G. (2000). *Evaluation : an integrated framework for understanding, guiding, and improving policies and programs*. San Francisco: Jossey-Bass.

Martin, M., & McIntyre, L. C. (1994). *Readings in the philosophy of social science*. Cambridge, Mass.: MIT Press.

McHugo, G. J., Drake, R. E., Brunette, M. F., Xie, H., Essock, S. M., & Green, A. I. (2006). Methodological issues in research on interventions for co-occurring disorders. *Schizophrenia Bulletin*, *32*, 655665.

McIntyre, L. C. (1994). Complexity and social scientific laws. In M. Martin & L. C. McIntyre (Eds.), *Readings in the philosophy of social science.* Cambridge, Mass.: MIT Press.

Minkoff, K. (2001). Developing standards of care for individuals with co-occurring psychiatric and substance use disorders. *Psychiatric Services*, *52*(5), 597599.

Mitchell, M. N. (2012). *Interpreting and visualizing regression models using stata.* College Station, Tex.: Stata Press.

Norcross, J. C., & Wampold, B. E. (2011, February). What works for whom: Tailoring psychotherapy to the person. *Journal of Clinical Psychology*, *67*(2), 127–132.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation.* London; Thousand Oaks, Calif.: Sage.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using stata vol. 1, continuous responses.* College Station, TX: Stata Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models : applications and data analysis methods* (Second edition ed.). Thousand Oaks, CA: Sage Publications.

Ripley, B. D. (1996). *Pattern recognition and neural networks.* New York: Cambridge University Press.

Roffman, J. L. (2011, June). Introduction. *Harvard Review of Psychiatry*, *19*(3), 99–101.

Rosenbaum, J. F., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.

Rossi, P. H., & Freeman, H. E. (1985). *Evaluation : a systematic approach.* Beverly Hills: Sage.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (2005, March). Causal inference using potential outcomes. *Journal of the American Statistical Association*, *100*(469), 322–331.

Scriven, M. (1956). A possible distinction between traditional scientific disciplines

and the study of human behavior. In *Minnesota studies in the philosophy of science* (Vol. 1, pp. 330–339). Minneapolis, MN: University of Minnesota Press.

Scriven, M. (1964). Views of human nature. In T. Wann (Ed.), *Behaviorism and phenomenology: Contrasting bases for modern psychology* (pp. 163–190). Chicago: University of Chicago Press.

Scriven, M. (2008). A summative evaluation of RCT methodology: & an alternative approach to causal research. *JOurnal of Multidisciplinary Evaluation*, *5*(9), 11–24.

Shadish, W. R. (2011, June). The truth about validity. *New Directions for Evaluation*, *2011*(130), 107–117.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation : theories of practice*. Newbury Park, Calif.: Sage Publications.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis : modeling change and event occurrence*. New York: Oxford University Press.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis*. Thousand Oaks, CA: Sage Publications.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348.

Substance Abuse and Mental Health Services Administration. (2011). *Join the voices for recovery: Recovery benefits everyone.* Washington, DC: Substance Abuse and Mental Health Services Administration.

Swarbrick, M. (2006). A wellness approach. *Psychiatric Rehabilitation Journal*, *29*(4), 311314.

Swayne, D. F., Cook, D., & Buja, A. (1998). XGobi: interactive dynamic data vi-

sualization in the x window system. *Journal of Computational and Graphical Statistics*, *7*(1), 113–130.

Tobin, T., Lang, D., & Huntington, N. (2012). *Assertive adolescent and family treatment cross-site evaluation final report.* Sudbury, MA: Advocates for Human Potential.

Twisk, J., & Proper, K. (2004, March). Evaluation of the results of a randomized controlled trial: how to define changes between baseline and follow-up. *Journal of Clinical Epidemiology*, *57*(3), 223–228.

U. S. Department of Housing and Urban Development. (2011). *The 2010 annual homeless assessment report to congress.* Washington, DC: U.S. Department of Housing and Urban Development.

Uher, R. (2011, June). Genes, environment, and individual differences in responding to treatment for depression. *Harvard Review of Psychiatry*, *19*(3), 109–124.

United States Interagency Council on Homelessness. (2010). *Opening doors: Federal strategic plan to prevent and end homelessness.* Washington, DC: United States Interagency Council on Homelessness.

Wikipedia. (2012). *Hasse diagram.*

Witten, I. H., & Frank, E. (2005). *Data mining : practical machine learning tools and techniques.* San Francisco: Kaufmann.